Universidade de Aveiro
Ano 2016

Departamento de Matemática

RODNEY CARVALHO AFONSO DE SOUSA

ANÁLISE ESTATÍSTICA DE DADOS COMPOSICIONAIS

Dissertação apresentada à Universidade de Aveiro para cumprimento dos requisitos necessários à obtenção do grau de mestre em Matemática e Aplicações, realizada sob a orientação científica da Doutora Adelaide Valente Freitas, Professora Auxiliar do Departamento de Matemática da Universidade de Aveiro.

Apoio financeiro da Fundação Calouste Gulbenkian, no âmbito do Programa de Bolsas de Pós-Graduação para Estudantes Africanos de Língua Oficial Portuguesa e Timor-Leste

O júri

Presidente

Professor Doutor Pedro Filipe Pessoa Macedo

Professor Auxiliar do Departamento de Matemática da Universidade de Aveiro

Professora Doutora Adelaide de Fátima Baptista Valente Freitas

Professora Auxiliar do Departamento de Matemática da Universidade de Aveiro

Professora Doutora Suzana Luísa de Custódia Machado Mendes

Professora Adjunta do Instituto Politécnico de Leiria - Escola Superior de Turismo e Tecnologia do Mar

agradecimentos

Para a realização desta dissertação diversas pessoas e entidades tiveram diferentes níveis de participação. Embora não seja possível enumerá-las todas, gostaria de expressar um agradecimento especial àquelas que contribuíram de forma mais direta e decisiva para o cumprimento deste objetivo.

Primeiramente, quero agradecer à Fundação Calouste Gulbenkian pelo financiamento do curso e acompanhamento do meu percurso académico nos últimos dois anos, sem os quais não me seria possível frequentar um programa de Mestrado em Matemática.

Um agradecimento a todos os professores do curso de Mestrado em Matemática e Aplicações da Universidade de Aveiro, pelo acolhimento e esclarecedoras lições ministradas nas aulas.

Agradeço aos meus antigos professores do curso de Licenciatura em Matemática do ex-Instituto Superior Politécnico de S. Tomé e Príncipe pelo incentivo e acompanhamento prestados no sentido de frequentar um curso de Pós-Graduação em Matemática.

Em especial, agradeço à Professora Doutora Adelaide Valente de Freitas, a orientadora deste trabalho, pelo interesse que me despertou nos temas abordados na disciplina de Análise Multivariada, e pela motivação, flexibilidade e disponibilidade e esclarecimentos prestados durante a realização desta dissertação.

Aos colegas e amigos que me acolheram neste país: Muito obrigado!

palavras-chave

dados composicionais, geometria de Aitchison, transformações log-razões, espaço dos codões, biplot.

resumo

Dados composicionais são dados multivariados em que cada unidade amostral corresponde a um vetor cujas componentes são números reais estritamente positivos, que representam proporções de um todo, e contêm apenas informação relativa, presente nas razões entre as suas componentes. Esse vetor está sujeito à restrição da soma das componentes ser igual à uma constante.

Podemos encontrar dados composicionais em muitos campos científicos, sendo que esses dados geralmente aparecem na forma de proporções, percentagens, concentrações, frequências absolutas ou relativas. Do ponto de vista geométrico, os dados composicionais pertencem a um subespaço real chamado simplex, sobre o qual se define uma geometria, chamada Geometria de Aitchison. Atualmente, a análise de dados composicionais baseia-se na análise estatística de log-razões (*logratios*) entre componentes (ou partes) das composições.

Neste trabalho, aplicamos técnicas exploratórias de dados composicionais na análise de um conjunto de dados do espaço dos codões referentes às regiões codificantes do ADN de 31 espécies distribuídas entre os cinco reinos de seres vivos: 16 animais, 4 plantas, 5 bactérias, 3 fungos e 3 protozoários. A principal ferramenta de análise utilizada é o biplot, que consiste numa representação gráfica que nos permite a visualização simultânea dos padrões existentes nas observações e nas variáveis de um conjunto de dados multivariado.

keywords

compositional data, Aitchison geometry, logratio transformations, codon space, biplot.

abstract

Compositional data are multivariate data consist of vectors of positive values summing to unit. They represent parts of a whole and contain only information presents in the ratios of its components.

We can find compositional data in many scientific areas. This kind of data usually appear as proportions, percentages, concentrations, absolute or relative frequencies. From a geometrical point of view, compositional data belong to a real subspace called simplex, where there is defined a specific geometry, called Aitchison geometry. Currently, the compositional data analysis is based on statistical analysis of log-ratios between components of the compositional vector.

In this work, we have used exploratory techniques of compositional data analysis to investigate patterns in a data set of the codon space concerning coding regions of DNA of 31 species distributed among the five kingdoms of living: 16 animals, 4 plants, 5 bacteria, fungi and 3 protozoa. The codon space is formed by the relative frequency of the four nucleotides in the three codon positions. The main analysis tool used is the biplot which is a graphical representation that allows the simultaneous visualization of patterns for the observations and variables of multivariate data.

Índice

LISTA DE FIGURAS

LISTA DE TABELAS

ABREVIATURAS

A	Nucleótido Adenina
ACP	Análise de Componentes Principais
ADN	Ácido desoxirribonucleico
ARN	Ácido ribonucleico
C	Nucleótido Citosina
G	Nucleótido Guanina
PBS	Partição Binária Sequencial
SVD	Decomposição em valores singulares (*Singular Value Decomposition*)
T	Nucleótido Timina
U	Nucleótido Uracilo

CAPÍTULO 1

INTRODUÇÃO

1.1. Noção de dados composicionais

Um vetor $\mathbf{x} = (x_1, x_2, \dots, x_D)$ é uma composição de D partes se todas as suas componentes são números reais estritamente positivos, que representam proporções de um todo, e contêm apenas informação relativa, presente nas razões entre as suas componentes. Esse vetor está sujeito à restrição de soma das componentes ser igual à uma constante, ou seja,

$$x_1 + x_2 + \cdots + x_D = k, \tag{1.1}$$

sendo $k > 0$ um número real. Geralmente temos $k = 1$ nos casos em que os dados forem medidos ou transformados para partes por unidades (ou proporções), ou $k = 100$ para medições feitas em percentagens (Pawlowsky-Glahn *et al.*, 2015).

■

Um conjunto de vetores D −dimensionais de observações com as características acima referidas é designado por dados composicionais *(compositional data)*. Podemos encontrar dados composicionais em muitos campos científicos, sendo que esses dados geralmente aparecem na forma proporções, percentagens, concentrações, frequências absolutas ou relativas. Visto que proporções são expressas em números reais, podemos ser tentados a interpretar ou analisar dados composicionais através da aplicação das tradicionais técnicas destinadas a dados multivariados reais. Essa prática pode levar-nos a paradoxos e/ou resultados sem significado no contexto do problema em estudo. Tal problemática tem sido abordada ao longo do tempo em áreas como Geologia, Biologia e Química (Pawlowsky-Glahn *et al*, 2015). Um dos primeiros exemplos vem do campo da morfologia biológica e é da autoria de um dos fundadores da Estatística moderna: Karl Pearson (1897). Em Geologia, o estudo de dados composicionais foi particularmente intenso entre 1960 e 1970. Porém, a primeira proposta metodológica consistente de análise de dados composicionais só chegou nos anos 1980, com os trabalhos de John Aitchison (1982, 1986). O principal aspecto da abordagem apresentada por Aitchison é a análise estatística de log-razões *(logratios)* entre as componentes de um vetor composicional e o estabelecimento dos princípios de uma análise de dados composicionais (Pawlowsky-Glahn *et al,* 2011).

Considerando que as composições fornecem apenas informação relativa entre as componentes, Aitchison (1986) concluiu que toda a análise das partes que compõem um todo poderia ser realizada em termos de razões das partes da composição. E, dado que a transformação log-razão é uma correspondência biunívoca em $\mathbb{R}$ e o tratamento matemático de um quociente é mais simples em termo de seu logaritmo, John Aitchison propôs metodologias baseadas em vários tipos de transformações log-razões. Essas transformações permitiram a aplicação de procedimentos da Análise Multivariada sobre os dados transformados traduzindo, de seguida, as conclusões extraídas em termos de dados originais (Pawlowsky-Glahn *et al*, 2015).

Apesar das vantagens oferecidas por técnicas baseadas em transformações log-razões na análise de dados composicionais, elas não tiveram o sucesso que se esperava no seio dos estatísticos. Tal facto talvez seja devido à tendência habitual de interpretar e analisar de resultados em termos absolutos e, consequentemente, a uma menor fluidez no raciocínio numa perspetiva relativa, o qual envolve pensar

em termos de razões. Assim, muitos investigadores têm continuado a aplicar os tradicionais métodos de Análise Estatística Multivariada aos dados composicionais, sem ter em conta o caráter composicional de seus dados. Na década de 2000 foram publicadas várias contribuições que permitiram uma melhor abordagem sistemática dos métodos propostos por John Aitchison (por exemplo, Pawlowsky-Glahn *et al,* 2001, 2003, Aitchison *et al*, 2002, 2005; Filzmoser *et al,* 2009). Atualmente, a análise de Dados composicionais pode ser basicamente descrita por três etapas: a representação de dados em coordenadas log-razões; uso de técnicas de análise estatística multivariada sobre os dados em coordenadas log-razões transformadas; e a interpretação dos resultados no contexto tanto das coordenadas transformadas como das coordenadas originais.

1.2. Motivação para o tema

Nesta seção apresentaremos alguns conceitos de Biologia Molecular necessárias para a compreensão dos objetivos do presente trabalho, nomeadamente sobre a estrutura primária do ácido desoxirribonucleico (ADN).

Existem milhões de espécies de seres vivos, tendo cada espécie caraterísticas funcionais e comportamentais próprias, que podem ser agrupadas em cinco reinos:

i. Monera – formado por seres unicelulares e procariotas (i.e., seres cujas células não possuem núcleo organizado). Fazem parte deste reino as bactérias e as algas azuis;
ii. Protista – formado por seres unicelulares e eucariotas (i.e., seres cujas células possuem um núcleo organizado). Fazem parte deste reino os protozoários e as algas inferiores;
iii. Fungo – formado por seres eucariotas uni ou pluricelulares, com parede celular formada por quitina. Fazem parte deste reino os fungos e os líquenes;
iv. Planta – formado pelos seres pluricelulares que possuem células revestidas por uma membrana de celulose e que são autótrofos (capazes de produzir a própria energia). Fazem parte deste reino os vegetais e as demais plantas;
v. Animal – formado por organismos pluricelulares e heterótrofos.

No entanto, quando analisamos esses organismos ao nível celular e molecular, observamos que estão organizadas de forma única na sua estrutura básica. A informação necessária para a formação de um novo organismo de cada espécie está contida no ADN. Esta informação genética é transferida de célula para célula e de pais para filhos. Assim, estudos genéticos objetivam compreender a forma como essas informações são transferidas e como elas podem ser modificadas (mutações), dando origem a diferentes organismos e espécies (Insana, 2003).

De uma forma simplificada, o ADN é representado através das quatro bases azotadas dos nucleótidos que são: a Adenina (A), a Citosina (C), a Guanina (G) e a Timina (T). Cada base azotada, juntamente com o ácido fosfórico e um açúcar, forma um nucleótido diferente. Por isso, muitas vezes (assim como neste trabalho), identificamos um nucleótido em termos de sua base azotada. As bases A e G são chamadas de purinas, enquanto T e C são chamadas de bases pirimidinas. Em termos matemáticos, uma sequência de ADN consiste numa sucessão das quatro bases (A, C, G, T), que constituem o alfabeto genético, e na qual está codificada toda informação sobre a estrutura e funções do organismo.

A molécula de ADN tem a estrutura de uma hélice dupla, em torno de um eixo central, onde A forma par com T e C forma par com G. A sequência de nucleótidos numa fita da hélice determina completamente a molécula de ADN. Nos anos 1950, Chargaff (1951) descobriu que a quantidade total

de nucleótidos da base pirimidina é sempre igual à quantidade total de nucleótidos da base purina (i.e., em termos de cardinalidade, C+T=A+G, sendo A=T e C=G). Contudo, a quantidade de A+T nem sempre é igual à de C+G.

Algumas sequências particulares de nucleótidos no ADN constituem unidades hereditárias, chamadas de genes, as quais determinam a produção de proteínas. Essas sequências constituem a parte codificante do ADN. Quando um gene se expressa, sua informação é primeiramente copiada para o ácido ribonucleico (ARN) que, em seguida, realiza a síntese de proteínas. As bases que formam o ARN são semelhantes às que formam o ADN, exceto no facto de que o nucleótido T é substituído pelo uracilo (U). Enquanto o ADN e o ARN possuem apenas 4 bases diferentes, as proteínas são constituídas por 20 unidades proteicas designadas de aminoácidos. Cada aminoácido é codificado por uma sequência de três nucleótidos. O código genético é lido em grupo de três bases, sendo cada grupo designado por codão. Um codão pode corresponder a um aminoácido numa proteína, ou a um codão de terminação da síntese de proteínas (*stop codon*). Embora existam apenas 20 aminoácidos conhecidos, o número de permutações possíveis das quatro bases de ADN são $4^3 = 64$ codões, pelo que existem codões que codificam o mesmo aminoácido (*synonymous codons*). Por exemplo, no código genético standard, existem 2 aminoácidos que são codificados por um só codão, 9 aminoácidos que são codificados (cada um) por dois codões, 5 aminoácidos que são codificados (cada um) por quatro codões, 1 aminoácido que é codificado por três codões, 3 aminoácidos que são codificados (cada um) por seis codões, e os três codões restantes correspondem aos codões de terminação. (Insana, 2003).

Um genoma consiste em toda a informação hereditária de um organismo, a qual está codificada no seu ADN, incluindo tanto os genes como as sequências não-codificantes (os chamados intrões).

Uma das mais básicas análises estatísticas realizadas sobre sequências de ADN de um conjunto de várias espécies corresponde à análise da distribuição das quatro bases no genoma ou nas sequências codificantes dessas espécies. Diversos estudos exploratórios revelam que as quatro bases têm distribuições diferentes (Takeuchi *et* al, 2003; Weir, 1996). Na Tabela 1.1 estão representadas a distribuição das quatro bases para cada uma das 31 espécies consideradas nesta dissertação. Podemos observar que as frequências das bases variam, quer quando comparamos as proporções de bases em sequência de um mesmo organismo, quer quando analisamos a proporção de uma dada base em sequências de organismos diferentes. Na Tabela A.2 em Anexos encontram-se as frequências absolutas das quatro bases em cada uma das três posições do codão, que deu origem à construção da Tabela 1.1.

Definição 1.1 (Espaço dos codões)

O espaço dos codões (*codon space*) é um espaço 12-dimensional, em que cada vetor contém as frequências dos quatro nucleótidos para cada uma das três posições do codão.

Assim, cada indivíduo (unidade amostral) do espaço dos codões corresponde a uma espécie, e é descrita por um vetor de 12 componentes $\mathbf{x} = (x_1, x_2, \dots, x_{12})$, sendo que as primeiras quatro componentes correspondem ao número de ocorrência dos nucleótidos A, C, G e T na primeira posição dos codões dessa espécie, as quatro seguintes correspondem ao número de ocorrência dos nucleótidos A, C, G e T na segunda posição dos codões, e as últimas quatro componentes correspondem ao número de ocorrência dos nucleótidos A, C, G e T na terceira posição dos codões. Considerando que cada posição de um codão só pode ser constituída por um dos quatro nucleótidos, significa que a contagem das quatro primeiras componentes determina o número h total dos codões

Tabela 1.1. Composições das bases das sequências codificantes do ADN das 31 espécies em estudo. (Consideramos a designação abreviada das espécies: nomes completos na Tabela A.1, em Anexos)

	Bases				
Espécies	**A**	**C**	**G**	**T**	**Total**
Bt	0.25	0.27	0.27	0.21	1,00
Cf	0.24	0.27	0.28	0.21	1,00
Eq	0.28	0.25	0.25	0.21	1,00
Gg	0.26	0.25	0.26	0.22	1,00
Dm	0.26	0.27	0.27	0.21	1,00
Um	0.26	0.25	0.26	0.22	1,00
Ay	0.29	0.20	0.24	0.27	1,00
Os	0.24	0.26	0.29	0.21	1,00
Po	0.29	0.20	0.23	0.28	1,00
Vv	0.28	0.20	0.23	0.28	1,00
Ba	0.35	0.15	0.21	0.29	1,00
Ec	0.24	0.25	0.27	0.24	1,00
Sa	0.36	0.15	0.19	0.30	1,00
St	0.31	0.19	0.22	0.29	1,00
Sm	0.31	0.17	0.21	0.31	1,00
Pl	0.45	0.10	0.14	0.31	1,00
Dd	0.41	0.14	0.14	0.32	1,00
Lm	0.19	0.31	0.31	0.18	1,00
Nc	0.24	0.29	0.27	0.20	1,00
SC	0.33	0.19	0.20	0.28	1,00
Sp	0.32	0.19	0.19	0.30	1,00
Ce	0.30	0.21	0.21	0.28	1,00
Dr	0.28	0.24	0.25	0.23	1,00
Hs	0.25	0.26	0.27	0.21	1,00
Mm	0.26	0.26	0.27	0.22	1,00
Pt	0.26	0.25	0.28	0.21	1,00
Rn	0.26	0.26	0.26	0.22	1,00
Ao	0.25	0.27	0.26	0.22	1,00
Fu	0.25	0.27	0.27	0.21	1,00
Xt	0.29	0.24	0.24	0.23	1,00
Am	0.34	0.18	0.21	0.27	1,00

no genoma de uma dada espécie. O mesmo se verifica para a contagem das quatro componentes centrais, bem como das quatro últimas componentes do vetor. Para cada espécie tem-se

1ª Posição				2ª Posição				3ª Posição				
A1	C1	G1	T1	A2	C2	G2	T2	A3	C3	G3	T3	TOTAL
x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}	x_{12}	$3h$

com

$$x_1 + x_2 + x_3 + x_4 = x_5 + x_6 + x_7 + x_8 = x_9 + x_{10} + x_{11} + x_{12} = h, \qquad (1.2)$$

em que h representa o número total dos codões no genoma, que varia de espécie para espécie.

■

Embora o número total dos codões (h) num genoma varia de espécie para espécie, os dados do espaço dos codões é de natureza composicional, porque o incremento de uma parte implica a alteração das outras partes.

Uma análise estatística do espaço dos codões de 27 espécies foi realizada por Takeuchi *et al* (2003), onde aplicaram uma Análise de Componentes Principais (ACP). A ACP permitiu a classificação de sequências codificantes das espécies em três grupos evolutivos, a saber: *Archaeas, Bactérias e Eucariotas.* Esta separação de espécies em grupos evolutivos foi determinada pela segunda componente principal. A primeira componente carateriza as espécies em termos do conteúdo CG em oposição ao conteúdo AT.

Ao analisar a proporção de nucleótidos para cada uma das três posições Takeuchi *et al* (2003) verificaram que, relativamente às 27 espécies consideradas, a Guanina (G) favorece a primeira posição do codão enquanto a Timina (T) é a que menos aparece nesta posição, e a Adenina é a que menos aparece na terceira posição. Além disso, verificaram que as bases na terceira posição apresentam maiores valores de desvios-padrão. No entanto, no seu estudo, Takeuchi *et al* (2003) apenas analisaram os dados numa perspetiva absoluta, sem considerar a natureza composicional dos dados, conforme propomos realizar neste trabalho.

1.3. Objetivos e organização da dissertação

Neste trabalho, pretendemos utilizar técnicas exploratórias de dados composicionais com o objetivo de analisar um conjunto de dados do espaço dos codões. Este conjunto de dados é formado pelas frequências relativas das bases nas três posições dos codões de 31 espécies distribuídas entre os cinco reinos de seres vivos: 16 animais, 4 plantas, 5 bactérias, 3 fungos e 3 protozoários. Esses dados resultaram de uma recolha realizada em 2010 de sequências das zonas codificantes do ADN de 31 espécies obtidas do National Center for Biotechnology Information (NCBI)[1]. Cada ficheiro de dados, com a informação das sequencias dos codões de uma espécie, foi processamento no software Anaconda[2] com vista a contabilizar o número de cada um dos quatro possíveis nucleótidos(A,C,G,T) em cada uma das três posições possíveis dos codões. Os dados cedidos correspondem a essas contagens (ver a Tabela A.2, em Anexos).

A principal ferramenta de análise que utilizaremos é o biplot, que consiste num tipo representação gráfica que nos permite a visualização simultânea dos padrões existentes nas observações e nas variáveis de um conjunto de dados multivariados. Com o objetivo de complementar as conclusões que podem ser extraídas da análise na perspetiva absoluta e na perspetiva relativa, aplicaremos biplots sobre dados em coordenadas originais (dados brutos) e sobre dados em coordenadas log-razões transformadas mais utilizadas na análise de dados composicionais. Adicionalmente, para os dados em cada uma das coordenadas referidas, compararemos os resultados obtidos por meio de biplots

[1] ftp://ftp.ncbi.nlm.nih.gov/genomes/archive/old_genbank/Eukaryotes/
[2] http://bioinformatics.ua.pt/software/anaconda/

clássicos e biplots robustos, sendo que estes últimos permitem contornar a eventual distorção dos resultados por parte da presença de observações atípicas (*outliers*) no conjunto de dados (Filzmoser *et al*, 2009).

Para atingir os objetivos acima propostos, esta dissertação está organizada do seguinte modo:

No Capítulo 2 definiremos o espaço de resultados para dados composicionais, juntamente com a sua respetiva geometria proposta por Aitchison. Apresentaremos ainda os princípios da Análise Composicional e as três transformações log-razões mais aplicadas aos dados composicionais.

No Capítulo 3 apresentaremos duas técnicas usadas para redução da dimensão de dados composicionais. A primeira técnica apresentada é conhecida como fusão (*amalgamation*) e consiste na soma das partes de uma composição. A outra técnica é designada equilíbrio (*balances*) entre grupos de partes de uma composição e é obtida através de um processo conhecido por partição binária sequencial (PBS).

No Capítulo 4 apresentaremos algumas técnicas da análise exploratória de dados composicionais. Abordaremos as descrições numérica e gráfica. Na descrição numérica abordaremos os conceitos de centro e tabela de variação, que correspondem, respetivamente, às medidas de localização e de dispersão de dados composicionais. Em relação às ferramentas gráficas, utilizaremos diagramas ternários e os biplots, sendo que os diagramas ternários são usados para analisar o padrão de variação entre três partes de uma composição, enquanto os biplots são ferramentas gráficas para a análise simultânea de possíveis padrões formados pelas observações e pelas variáveis.

Por fim, no Capítulo 5, consideraremos um conjunto de dados do espaço dos codões, constituído pelas 31 espécies listadas na Tabela 1.1, para as quais exploraremos a variação absoluta e relativa das frequências dos nucleótidos, considerando diferentes situações (4 casos de estudos) usando técnicas de análise de dados composicionais abordadas nos Capítulos 2 a 4.

As aplicações práticas foram feitas com recurso ao software estatístico **R** (R Core Team, 2014). Para a importação dos dados de um ficheiro Excel recorremos ao pacote `RODBC` (Ripley *et al,* 2014). Para aplicação das técnicas de dados composicionais recorremos aos pacotes `compositions` (van den Boogaart *et al,* 2014) e `mvoutlier` (Filzmoser *et al*, 2015). Para determinação de estimativas robustas da matriz de variância-covariâncias e do vetor das médias recorremos ao pacote `rrcov` (Todorov *et al, 2*009). Os scripts encontram-se em Anexos.

CAPÍTULO 2

GEOMETRIA DE AITCHISON

2.1. Introdução

Nesta seção abordaremos alguns conceitos necessários para a compreensão do objetivo principal deste trabalho, principalmente as propriedades inerentes à estrutura de dados composicionais, e apresentaremos alguns exemplos que realçam a importância de se levar em conta a natureza composicional de dados em análises estatísticas.

Comecemos por notar que vetores com componentes positivas proporcionais representam a mesma composição, pois, a multiplicação de um vetor de componentes positivas por uma constante positiva não muda a razão entre as componentes. Isto sugere que composições podem ser vistas como classes de equivalência de vetores proporcionais, contendo a mesma informação.

Definição 2.1 (Composições como classes de equivalência)

Dois vetores $\mathbf{x}, \mathbf{y} \in \mathbb{R}_+^D$ $(x_i, y_i > 0, \forall\, i = 1, 2, \dots, D)$ são composicionalmente equivalentes se existe um escalar $\lambda \in \mathbb{R}_+$ tal que $\mathbf{x} = \lambda \cdot \mathbf{y}$, ou seja, as composições $\mathbf{x} = (x_1, x_2, \dots, x_D)$ e $\boldsymbol{y} = (\lambda x_1, \lambda x_2, \dots, \lambda x_D)$ contêm essencialmente a mesma informação relativa, $\forall\, \lambda \in \mathbb{R}_+$.

■

Qualquer vetor de uma classe de equivalência pode ser usado para representá-la. Deste modo, qualquer composição pode ser expressa em proporções utilizando-se um fator de escala apropriado. De modo a facilitar qualquer análise, convém selecionar um representante da classe de equivalência, pela normalização dos vetores, de modo que a soma das componentes iguale a uma dada constante k, que pode ser $1, 100, 1000, 10^6$ ou qualquer outra constante positiva. Esta seleção pode ser formalizada por uma operação designada por fecho (*Closure*).

Definição 2.2 (Fecho de uma composição)

Seja $\mathbf{x} = (x_1, x_2, \dots, x_D)$ um vetor de componentes reais e estritamente positivos e $k \in \mathbb{R}_+$ uma contante. O fecho de $\mathbf{x}$ a k é uma operação que transforma uma composição $\mathbf{x}$ noutra composição equivalente, $C(\mathbf{x})$, dada por:

$$C(\mathbf{x}) = \left(\frac{k \cdot x_1}{\sum_{i=1}^{D} x_i}, \frac{k \cdot x_2}{\sum_{i=1}^{D} x_i}, \dots, \frac{k \cdot x_D}{\sum_{i=1}^{D} x_i} \right). \tag{2.1}$$

O resultado do fecho é uma reestruturação do vetor inicial, de modo que a soma de suas componentes seja igual a k. Assim, pode-se dizer que dois vetores $\mathbf{x}$ e $\mathbf{y}$ em $\mathbb{R}_+^D$ são equivalentes se, para qualquer constante $k \in \mathbb{R}_+$, tem-se $C(\mathbf{x}) = C(\mathbf{y})$ (Pawlowsky-Glahn *et al.*, 2015).

■

Exemplo 2.1. Fecho de uma composição

Consideremos a composição referente às frequências absolutas dos quatro nucleótidos na parte codificante do ADN da espécie Hs, dada por $\mathbf{x} = (13981961, 14606833, 15036255, 11595288)$.

Como $x_1 + x_2 + x_3 + x_4 = 55220337$, então o fecho da composição $\mathbf{x}$ a $k = 1$ é dado por

$$\mathbf{y} = C(\mathbf{x}) = \left(\frac{1 \times 13981961}{55220337}, \frac{1 \times 14606833}{55220337}, \frac{1 \times 15036255}{55220337}, \frac{1 \times 11595288}{55220337}\right)$$
$$= (0.253, 0.265, 0.272, 0.210),$$

em que a composição resultante $\boldsymbol{y}$ satisfaz $y_1 + y_2 + y_3 + y_4 = 1$. Além disso, $\mathbf{x}$ e $\mathbf{y}$ são composições equivalentes, pois, podemos escrever $\mathbf{x} = 55220337\mathbf{y}$**.**

■

A restrição de soma constante confere a dados composicionais caraterísticas particulares, tornando-se necessário definir o espaço dos dados composicionais como um espaço que atenda a essas particularidades.

Definição 2.3 (Conjunto Simplex)

O espaço amostral de dados composicionais de D partes, designado por D −simplex, e denotado por S^D, é definido por

$$S^D = \left\{x = (x_1, x_2, \dots, x_D) \in \mathbb{R}_+^D : \sum_{i=1}^{D} x_i = k\right\}. \quad (2.2)$$

■

Embora uma composição seja uma classe de equivalência, os representantes dessa classe no simplex também são chamados de composições. As componentes de um vetor em S^D são chamadas de partes para salientar o seu caráter composicional.

Muitas vezes, no estudo de um conjunto de dados composicionais, o interesse pode estar apenas em algumas partes de uma composição.

Definição 2.4 (Subcomposição)

Dada uma composição $\mathbf{x}$ e uma seleção de índices $S = \{i_1, i_2, \dots, i_s\}$, uma subcomposição $\mathbf{x}_s$**,** com s partes, é obtida pela aplicação da operação de fecho ao subvetor $(x_{i_1}, x_{i_2}, \dots, x_{i_s})$ de $\mathbf{x}$**.** O conjunto de índices S indica as partes selecionadas para a subcomposição.

■

Exemplo 2.2. Uma subcomposição para o espaço dos codões

Consideremos no espaço dos codões a espécie Hs. No espaço dos codões cada composição é um vetor de 12 partes $\mathbf{x} = (x_1, x_2, \dots, x_{12})$, conforme definido no Capítulo 1. Se pretendemos estudar apenas a composição do nucleótido A nas três posições do codão, deverá ser considerada a subcomposição $\mathbf{y}$ de $\mathbf{x}$ dada por

$$\mathbf{y} = C(x_1, x_5, x_9),$$

que no caso da espécie Hs ficaria

$$\mathbf{y} = \left(\frac{0.2618}{0.2618 + 0.2957 + 0.2027}, \frac{0.2957}{0.2618 + 0.2957 + 0.2027}, \frac{0.2027}{0.2618 + 0.2957 + 0.2027}\right)$$
$$= (0.3444, 0.3890, 0.2666).$$

■

Visto que proporções são expressos em números reais, somos tentados a interpretar ou a analisar dados composicionais como dados multivariados reais sem ter em conta as suas caraterísticas especiais, nomeadamente, de pertencer ao simplex. Essa prática pode levar-nos a paradoxos e/ou resultados sem significados, ou ainda a interpretações erradas dos resultados no contexto do problema em estudo (Pawlowsky-Glahn *et al*, 2015). Vejamos alguns exemplos que ilustram esses paradoxos.

2.1.1. O problema da correlação espúria

Em 1897 Karl Pearson publicou um artigo[3], cujo título incluía a expressão: *"Em uma forma de correlação espúria…"* com o qual pretendia alertar a comunidade científica da época sobre alguns problemas relacionados com a análise estatística de dados composicionais (Aitchison, 1986, pág 48). Referindo ao mesmo artigo, Aitchison citou as palavras de Karl Pearson da seguinte forma: «*Cuidado com as tentativas de interpretar as correlações entre índices cujos numeradores e denominadores contêm partes em comum*» (Aitchison, 2005, pág. 13). Isto geralmente acontece quando se lida com dados composicionais.

Na realidade, apesar do alerta emitido por karl Pearson, continuaram a aparecer trabalhos onde se calculava a correlação de Pearson de componentes de dados composicionais considerando a usual interpretação para dados multivariados em $\mathbb{R}^n$ sem restrições.

A principal questão relacionada com a análise de dados composicionais por meio de métodos da usual Estatística Multivariada prendia-se com a impossibilidade de se interpretar os coeficientes de correlação de Pearson entre as componentes dos dados originais e ficou conhecido na literatura como o problema de correlação espúria (*Spurious correlations*). O problema de correlação espúria refere-se à existência de uma relação estatística entre duas ou mais variáveis, mas onde não existe nenhuma explicação lógica ou significado teórico. Tal ocorre com frequência quando lidamos com dados em que a soma das componentes é constante.

Por exemplo, para uma composição de D partes $\mathbf{x} = (x_1, x_2, \dots, x_D)$ sujeita à restrição da soma das componentes ser igual à unidade, isto é, $x_1 + x_2 + \cdots + x_D = 1$, resulta que

$$cov(x_1, x_1 + x_2 + \cdots + x_D) = 0,$$

e portanto,

$$cov(x_1, x_2) + cov(x_1, x_3) + \cdots + cov(x_1, x_D) = -var(x_1). \qquad (2.3)$$

Consequentemente, o segundo membro da equação (2.3) é sempre negativo, exceto para o caso em que a primeira componente x_1 é constante. Assim sendo, pelo menos uma das covariâncias do primeiro membro deve ser negativa ou, de modo equivalente, deve haver pelo menos um elemento negativo na primeira linha da matriz de covariâncias dos dados originais. Esse efeito é chamado de viés negativo (*negative bias*) e induz à existência de correlações espúrias entre as variáveis. O mesmo viés negativo deve ocorrer em outras linhas, afetando pelo menos D dos elementos da matriz de covariâncias dos dados originais (Aitchison, 2005). Assim, a aplicação da análise de correlação usual

[3] *Mathematical contributions to the theory of evolution. On a formo f spurious correlation which may arise when indices are used in the measurement of organism*

para esse tipo de dados pode conduzir a resultados que não permitem uma correta interpretação da relação entre as variáveis.

Exemplo 2.3. Correlação espúria *(Adaptado de Aitchison (2005), pág. 21)*

Consideremos dois cientistas, A e B, interessados em amostras de um solo que tenham sido subdivididas em três grupos. Para cada grupo da amostra, o cientista A regista uma composição de quatro partes (animal, vegetal, mineral e água); o cientista B primeiramente seca cada grupo, sem registar o teor de água, e obtém uma composição de três partes (animal, vegetal e mineral). Assuma-se, por simplicidade, que os grupos em cada um dos casos eram idênticos, e que os cientistas foram precisos nas suas medições. Representemos cada uma das partes da composição obtida pelo cientista A por x_1, x_2, x_3 e x_4, respetivamente, animal, vegetal, mineral e água; e as partes da composição obtida pelo cientista B por x_1', x_2' e x_3' respetivamente, animal, vegetal e mineral. Tendo em conta o significado das partes das duas composições, é evidente que a composição obtida por B é uma subcomposição de A. Assim, as conclusões chegadas pelos cientistas A e B, na análise de partes em comum, deverão estar de acordo. Os dados estão na Tabela 2.1.

Tabela 2.1. Amostras de composição do solo registadas pelos cientistas A e B.

	Cientista A				Cientista B		
Amostras	x_1	x_2	x_3	x_4	x'_1	x'_2	x'_3
1	0,1	0,2	0,1	0,6	0,25	0,50	0,25
2	0,2	0,1	0,2	0,5	0,40	0,20	0,40
3	0,3	0,3	0,2	0,2	0,43	0,43	0,14

Para avaliar a relação entre duas partes, x_i e x_j, os dois cientistas podem determinar o coeficiente de correlação de Pearson entre elas, dado por

$$\rho = \frac{cov(x_i, x_j)}{\sqrt{var(x_i) \cdot var(x_j)}}. \tag{2.4}$$

Uma interpretação para o valor do coeficiente de correlação está na Tabela 2.2 (Filho *et al*, 2009).

Tabela 2.2. Intervalos de referência para interpretação do coeficiente de correlação

Valor de ρ	**Interpretação**
$0 \leq \lvert\rho\rvert < 0{,}3$	Correlação desprezível
$0{,}3 \leq \lvert\rho\rvert < 0{,}3$	Correlação fraca
$0{,}4 \leq \lvert\rho\rvert < 0{,}7$	Correlação moderada
$0{,}7 \leq \lvert\rho\rvert \leq 1{,}0$	Correlação forte

A soma das partes de cada amostra registadas pelos cientistas A e B é sempre igual à unidade. Por outro lado, como nenhuma das partes é constante sabemos que vai ocorrer o efeito de viés negativo na matriz de covariâncias (Tabela 2.3) e consequentemente ocorrerá correlações espúrias na matriz de correlações (Tabela 2.4).

Na Tabela 2.3, observamos que o viés negativo ocorre nas três primeiras linhas da matriz de covariâncias dos dados registados pelo cientista A, pois verifica-se que a soma dos elementos de cada uma dessas linhas é igual a 0 e, consequentemente, verifica-se a igualdade $\sum_j cov(x_i, x_j) = -var(x_i)\,, i = 1, 2, 3, 4,\ com\ i \neq j$. O mesmo viés negativo ocorre em todas as linhas da matriz de covariâncias de dados registados pelo cientista B.

Tabela 2.3. Matrizes de covariâncias de amostras registadas pelo cientista A e pelo cientista B

	Cientista A					Cientista B		
Partes	x_1	x_2	x_3	x_4	Partes	x'_1	x'_2	x'_3
x_1	**0,010**	0,005	0,00	-0,015	x'_1	**0,009**	-0,008	-0,001
x_2	0,005	**0,010**	-0,005	-0,010	x'_2	-0,008	**0,024**	-0,016
x_3	0,00	-0,005	**0,003**	0,002	x'_3	-0,001	-0,016	**0,017**
x_4	-0,015	-0,010	0,0017	0,023				

Tabela 2.4. Matriz de correlações de amostras registadas pelo cientista A e pelo cientista B

	Cientista A					Cientista B		
Partes	x_1	x_2	x_3	x_4	Partes	x'_1	x'_2	x'_3
x_1	1,00	**0,50**	0,00	-0,98	x'_1	1,00	**-0,56**	-0,07
x_2		1,00	-0,87	-0,65	x'_2		1,00	-0,79
x_3			1,00	0,19	x'_3			1,00
x_4				1,00				

Ao analisar as correlações entre as partes das composições, de acordo com as matrizes de correlações representadas na Tabela 2.4, observamos que, para o cientista A, a correlação entre animal e vegetal é $corr(x_1, x_2) = 0{,}5$, o que sugere que existe uma moderada correlação positiva entre os conteúdos de animal e de vegetal no solo de onde foram recolhidas as amostras. No entanto, a correlação entre essas componentes, determinada pelo cientista B seria $corr(x'_1, x'_2) = -0{,}56$, sugerindo a existência de uma moderada correlação negativa entre os conteúdos de animal e vegetal no mesmo solo, o que representa uma grande inconsistência entre as conclusões chegadas por ambos. Esta inconsistência não nos permite interpretar a relação entre as partes envolvidas.

■

O problema de correlação espúria, conforme ilustrado no *Exemplo 2.3*. ocorre com frequência quando se analisa um conjunto de dados cuja soma das componentes é uma constante, ou um subconjunto de mesmas partes, cuja soma é também uma contante (neste caso, 1) (Pawlowsky-Glahn *et al*, 2015). Ao longo do século XX, problemas desse tipo receberam vários nomes, tais como problema da soma constante ou problema do fecho (Kucera *et al,* 1997), problema do viés negativo e dificuldade de correlação nula (Gallo *et al,* 2007). No entanto, não se registou nenhuma tentativa no sentido de se desenvolver técnicas de análise estatística que se ajustassem às caraterísticas particulares de dados composicionais. Na verdade, Aitchison (2003) refere que, perante um resultado inconsistente ou sem qualquer explicação teórica derivado da análise de dados composicionais, os analistas procuravam, essencialmente, verificar o que tinha saído errado na aplicação de técnicas usuais de análise multivariada a dados composicionais na esperança de ser possível aplicar algumas correções.

Uma metodologia adequada, e o estabelecimento de princípios lógicos necessários para a análise de dados composicionais e a natureza especial de seu espaço amostral começou a aparecer na década de 1980, com trabalhos de John Aitchison (1980, 1982, 1983, 1985), e culminou com uma monografia metodológica de sua autoria, intitulada *The Statistical Analysis of Compositional Data*, publicada em 1986.

2.1.2. Simplex como espaço vetorial

Aitchison (1986) introduziu duas operações puramente composicionais, conhecidas na literatura por perturbação (*perturbation*) e potenciação (*powering*), que permitem conferir ao simplex de D partes a estrutura de um espaço vetorial e, deste modo, definir bases, linha retas e outros operadores no simplex.

Definição 2.5 (Perturbação)
Consideremos duas composições $\mathbf{x}, \mathbf{y} \in S^D$. A perturbação de $\mathbf{x}$ por $\mathbf{y}$ é definida como a composição

$$\mathbf{x} \oplus \mathbf{y} = C(x_1 y_1, x_2 y_2, \dots, x_D y_D), \tag{2.5}$$

em que $C(\cdot)$ é a operação de fecho.

■

Quando ocorrem alterações de valores em algumas ou em todas as partes de uma composição estamos perante uma perturbação. Este tipo de processo ocorre com frequência na Química, quando por exemplo, as concentrações em partes por milhão (ppm) de peso são alterados para concentrações molares, pois, tal corresponde à multiplicação de cada componente pelo inverso do peso molar. Também na Geologia, quando consideramos uma composição de um sedimento, por exemplo de três partes $\mathbf{x} = (x_1, x_2, x_3)$, em que (x_1, x_2, x_3) refere aos teores de areia, silte e argila, respetivamente, que após um processo de erosão, é depositada uma composição $\mathbf{y}$ que representa proporções de cada uma das partes da composição $\mathbf{x}$. A composição resultante é dada por $\mathbf{x} \oplus \mathbf{y}$. Note que a operação de fecho em (2.5) garante que a composição resultante mantenha o seu caráter composicional (Pawlowsky-Glahn *et al,* 2006, 2011).

Devemos observar que para $\boldsymbol{n} = C(1, 1, \dots, 1) = (1/D, 1/D, \dots, 1/D)$, temos $\boldsymbol{n} \oplus \mathbf{x} = \mathbf{x}$. Assim, uma composição com todas as componentes iguais define o elemento neutro da perturbação. Em termos matemáticos, o par $(S^D, \oplus)$ forma um grupo comutativo[4], sendo $\boldsymbol{n} = \left(\frac{1}{D}, \frac{1}{D}, \dots, \frac{1}{D}\right)$ o elemento neutro da perturbação, e $\mathbf{y^{-1}} = C(y_1^{-1}, y_2^{-1}, \dots, y_D^{-1})$ a inversa de uma composição $\mathbf{y}$.

Assim, a perturbação de $\mathbf{x}$ pela inversa de $\mathbf{y}$ é dada por $\mathbf{x} \oplus \mathbf{y}^{-1}$ e denota-se por $\mathbf{x} \ominus \mathbf{y} = C\left(\frac{x_1}{y_1}, \frac{x_2}{y_2}, \dots, \frac{x_D}{y_D}\right)$.

Exemplo 2.4. Inversa da perturbação – Frequências de nucleótidos das espécies Hs e Mm

Consideremos as frequências dos nucleótidos A, C, G e T nas três posições dos codões das espécies Hs e Mm, conforme representados na Tabela 1.1 e representemos a composição referente a Hs por $\mathbf{x} = (0.25, 0.26, 0.27, 0.21)$ e a composição referente a Mm por $\mathbf{y} = (0.26, 0.26, 0.27, 0.22)$. Podemos medir a mudança ocorrida nas frequências de nucleótidos dessas duas espécies como a perturbação de $\mathbf{y}$ pela inversa de $\mathbf{x}$:

$$\begin{aligned}
\mathbf{y} \ominus \mathbf{x} &= C\left(\frac{0.25}{0.26}, \frac{0.26}{0.26}, \frac{0.27}{0.27}, \frac{0.21}{0.22}\right) \\
&= C(1.04, 1.00, 1.00, 1.048) \\
&= (1.04, 1.00, 1.00, 1.048)/(1.04 + 1.00 + 1.00 + 1.048)
\end{aligned}$$

[4] É uma estrutura algébrica, também chamada de grupo abeliano, que satisfaz: $a \oplus b = b \oplus a, \forall a, b \in S^D$

$$= \left(\frac{1.04}{4.09}, \frac{1.00}{4.09}, \frac{1.00}{4.09}, \frac{1.048}{4.09}\right)$$

$$= (0.26, 0.24, 0.24, 0.26)$$

O valor da perturbação é aproximadamente igual ao elemento neutro do simplex S^4, dado por $\boldsymbol{n} = (0.25, 0.25, 0.25, 0.25)$, o que significa que a diferença relativa entre as frequências de nucleótidos das espécies Hs e Mm é muito reduzida.

■

Definição 2.6 (Potenciação)

Consideremos uma composição $\mathbf{x} \in S^D$ e um escalar $\alpha \in \mathbb{R}$. A potenciação de $\mathbf{x}$ por α é uma composição designada por $\alpha \otimes \mathbf{x}$, e dada por

$$\alpha \otimes \mathbf{x} = C(x_1^\alpha, x_2^\alpha, \dots, x_D^\alpha). \tag{2.6}$$

■

A potenciação da composição $\mathbf{x}$ por α pode ser visto como a perturbação de $\mathbf{x}$ por si mesma α vezes. A potenciação, juntamente com as propriedades da perturbação, conferem ao simplex S^D a estrutura de um espaço vetorial. Assim, dado um ponto $\mathbf{x_0} \in S^D$ e um vetor $\mathbf{v} \in S^D$, definimos linha reta composicional com origem $\mathbf{x_0}$ e direção $\boldsymbol{v}$ pela seguinte equação:

$$\mathbf{x}(\alpha) = \boldsymbol{x_0} \oplus (\alpha \otimes \boldsymbol{v}), \quad \alpha \in \mathbb{R}. \tag{2.7}$$

Esta representação de uma linha reta no simplex S^D é importante na definição de modelos lineares básicos na análise de dados composicionais. Em particular, são usadas pra identificação de tendências na representação de dados no simplex (Pawlowsky-Glahn *et al*, 2006).

2.2. Princípios de análise composicional

2.2.1. Introdução

John Aitchison (1986) indicou três princípios sobre os quais se devem reger as técnicas adequadas de análise de dados composicionais. Ao definir esses princípios, aquele autor considerou que numa análise estatística de dados composicionais apenas as proporções das componentes contêm informações relevantes. Os três princípios são:

- Invariância de escala;
- Invariância de permutação;
- Coerência subcomposicional.

O significado de cada um desses princípios no contexto de análise estatística de dados composicionais serão ilustrados nas subseções que se seguem.

2.2.2. Invariância de escala

Princípio: *Quando um problema é composicional, devemos reconhecer que o valor absoluto das partes que compõem as amostras são irrelevantes, uma vez que composições equivalentes contêm essencialmente a mesma informação.*

Por exemplo, consideremos dois vetores $\mathbf{x} = (1.6, 2.4, 4.0)$ e $\mathbf{y} = (3.0, 4.5, 7.5)$ em $\mathbb{R}^3_+$, representando, respetivamente, pesos de três partes (a, b, c) de dois espécimes de uma rocha, de peso total 8 g e 15 g, respetivamente. Se pretendemos fazer análise composicional dessa rocha, devemos reconhecer que $\mathbf{x}$ e $\mathbf{y}$ representam a mesma composição pois são vetores equivalentes, onde a diferença de pesos a ser levada em consideração é dada pela relação de escala $\mathbf{y} = \frac{15}{8}\mathbf{x}$. Assim, um requisito fundamental na análise de dados composicionais é que uma função adequada deve ser tal que $f(\mathbf{y}) = f(\mathbf{x})$, sempre que $\mathbf{x}$ e $\mathbf{y}$ forem vetores equivalentes (Aitchison, 2005). Uma função com tal propriedade é chamada de função invariável quanto a escala (*scale invariant*).

A seguir, apresentamos uma definição formal de função invariante quanto à escala.

Definição 2.7 (Função invariante quanto à escala)
Seja f uma função definida em $\mathbb{R}^D_+$. Essa função é invariante quanto à escala se, para qualquer número real positivo $\lambda \in \mathbb{R}_+$ e para qualquer composição $\mathbf{x} \in S^D$, satisfaz $f(\lambda\mathbf{x}) = f(\mathbf{x})$, isto é, a imagem de vetores composicionalmente equivalentes por meio de f é sempre a mesma.

■

2.2.3. Invariância de permutação

Princípio: *As conclusões de uma análise composicional não deve depender da ordem das partes envolvidas.*

Por exemplo, em composições geológicas é muito frequente o registo de partes por ordem alfabética. Aplicando a análise composicional, a ordem das diferentes partes não deve desempenhar qualquer papel relevante.

2.2.4. Coerência subcomposicional

Princípio: *As análises sobre um conjunto de partes de uma composição não devem depender de outras partes não envolvidas, pelo que o estudo de uma subcomposição não pode conduzir a resultados contraditórios com os obtidos a partir da composição total.*

Egozcue *et al* (2007) resume o princípio de coerência subcomposicional a dois critérios que são:

- A distância medida entre duas composições completas deve ser maior ou igual à distância entre quaisquer de suas subcomposições. Este comportamento é chamado de dominância subcomposicional (*Subcomposicional dominance*)*;*
- Se eliminarmos uma parte "não-informativa" de nossos dados composicionais, os resultados da nossa análise não devem mudar.

Consequentemente, técnicas adequadas de análise de dados composicionais deve garantir que a seleção de uma subcomposição não altere a relação entre as partes, ou seja, visto que as proporções das partes constituem a única informação considerada, a análise deve manter-se invariável quando se usa as mesmas partes da composição e da subcomposição. Por exemplo, retomando o *Exemplo 2.3*, em que o cientista B analisa uma subcomposição da amostra analisada pelo cientista A, técnicas adequadas de análise deverão conduzir os dois analistas à mesma conclusão com relação às partes animal, vegetal e mineral.

■

Visando satisfazer os requisitos relativos aos princípios de análise composicional, John Aitchison sugeriu uma nova geometria para o simplex S^D. O desenvolvimento dos conceitos sugeridos por Aitchison (1986) deu origem à conhecida Geometria de Aitchison para o simplex. Visto que se trata de uma geometria euclidiana num espaço transformado, requer definições e métrica específicas.

2.3. Transformações de dados composicionais

2.3.1. Introdução

Na análise composicional apenas a informação relativa presente nas proporções das partes é relevante. Atualmente, esta análise baseia-se essencialmente na análise estatística de log-razões das partes, no espaço real, onde se pode aplicar técnicas usuais de análise multivariada que, após a conclusão das mesmas, podem ser traduzidas em termos dos dados originais (não transformadas). Esta metodologia de análise estatística para dados composicionais é conhecida na literatura como Análise de Log-razões (*Logratio Analysis*) (Aitchison, 2005; Pawlowsky-Glahn *et al,* 2011). Esta abordagem surgiu em resultado do reconhecimento da importância do princípio de invariância de escala, cuja aplicação prática exigia que se trabalhasse com razões entre as componentes, que anula a constante de escala. Considerando que a transformação log-razão é uma correspondência biunívoca em $\mathbb{R}$ e o tratamento matemático de um quociente é mais simples em termo de seu logaritmo, Aitchison propôs a adoção de uma técnica de transformação envolvendo logaritmos de razões das componentes (Aitchison, 2005).

Dada uma composição $\mathbf{x} = (x_1, x_2, \dots, x_D)$, podemos definir diversas transformações log-razões. A mais simples é aquela que relaciona duas partes, $ln(x_i/x_D),\ i = 1, 2, \dots, D-1$, sendo que poderia figurar no denominador qualquer uma das partes para além de x_D (Aitchison, 1986). Neste caso, uma transformação de log-razões das partes de $\mathbf{x}$ é uma composição $\mathbf{y}$ definida do seguinte modo:

$$\mathbf{y} = \left(ln\left(\frac{x_1}{x_D}\right), \dots, ln\left(\frac{x_{D-1}}{x_D}\right)\right).$$

Embora esta composição transformada $\boldsymbol{y}$ contenha apenas $D-1$ partes, a partir desta, também podemos recuperar a composição original $\mathbf{x}$ do seguinte modo:

$$(x_1, x_2, \dots, x_D) = (exp(y_1), exp(y_2), \dots, exp(y_{D-1}), 1)/(exp(y_1) + \cdots + exp(y_{D-1}) + 1).$$

Um conceito importante na análise de dados composicionais é o de log-contraste.

Definição 2.8 (Log-contraste)

Seja $\mathbf{x} = (x_1, x_2, \dots, x_D)$ uma composição. Um log-contraste de $\mathbf{x}$ é uma combinação log-linear definida do seguinte modo:

$$\boldsymbol{a}' log\mathbf{x} = \sum_{i=1}^{D} \alpha_i\, ln(x_i)\,, \tag{2.8}$$

com

$$\sum_{i=1}^{D} \alpha_i = 0. \tag{2.9}$$

■

Log-contrastes podem ser encarados como uma combinação linear no simplex, e gozam de propriedades interessantes para a análise de dados composicionais. Algumas dessas propriedades são as seguintes (Aitchison, 1986):

a) Log-contrastes são invariantes quanto à escala, pois

$$\begin{aligned}\boldsymbol{a}'log(k\mathbf{x}) &= \sum_{i=1}^{D} \alpha_i\, ln(kx_i) \\ &= \sum_{i=1}^{D} \alpha_i\, ln(x_i) + \sum_{i=1}^{D} \alpha_i\, ln(k) \\ &= \sum_{i=1}^{D} \alpha_i\, ln(x_i) + ln(k) \times \sum_{i=1}^{D} \alpha_i \\ &= \sum_{i=1}^{D} \alpha_i\, ln(x_i) + ln(k) \times 0 \\ &= \sum_{i=1}^{D} \alpha_i\, ln(x_i) = \boldsymbol{a}'\, ln(\mathbf{x}).\end{aligned}$$

b) A condição (2.9) garante que a combinação linear de log-razões entre partes de uma composição $\mathbf{x} \in S^D$ é um log-contraste, nomeadamente:

   (i) a transformação log-razões usando uma das partes fixa no denominador, por exemplo x_D,

$$\sum_{i=1}^{D-1} \alpha_i\, ln(x_i/x_D),$$

   ou matricialmente na forma $\boldsymbol{a}'ln(\mathbf{x}_{-D}/x_D)$, em que $\mathbf{x}_{-\boldsymbol{D}}$ representa a composição que se obtém pela remoção da componente D do vetor $\mathbf{x}$;

   (ii) a transformação log-razões alternativa, usando a média geométrica $g(\mathbf{x})$ no denominador

$$\sum_{i=1}^{D} \alpha_i ln(x_i/g(\mathbf{x}))$$

As transformações log-razões invariantes quanto à escala são log-contrastes (*logcontrasts*)

Definição 2.9 (**Log-contrastes ortogonais**)

Dois log-contrastes $\boldsymbol{a}'\log(\mathbf{x})$ e $\boldsymbol{b}'\log(\mathbf{x})$ são ortogonais se tivermos $\boldsymbol{a}'\boldsymbol{b} = 0$.

■

Geralmente, muitas dificuldades relacionadas com a análise de dados composicionais podem ser ultrapassadas pela análise de um log-contraste apropriado. A escolha do log-contraste depende do problema e da interpretação da composição. Representações adequadas e completas de uma composição através de um conjunto de log-contrastes foram propostas de modo que todas as informações da composição são convertidas para o conjunto de logaritmos de razões das partes da composição (Aitchison, 1986). A primeira proposta foi a transformação de log-razões aditiva (alr: $additive\ log-ratio$). Após perceber que esta transformação não era isométrica, no sentido

que as relações entre as distâncias no espaço transformado são alteradas, Aitchison introduziu a transformação de log-razões centradas (clr: $centered\ log-ratio$), baseada na média geométrica das partes das composições. Mais recentemente, Egozcue *et al* (2003) propôs a transformação de log-razões isométrica *(*ilr: $isometric\ log-ratio$) definida a partir de uma base ortonormal no simplex.

Nas subseções que se seguem analisaremos as três transformações acima referidas.

2.3.2. Transformação alr

Definição 2.10 (Transformação alr)

Seja $\mathbf{x}$ uma composição de D partes no simplex S^D. Chama-se transformação de log-razões aditivas de $\mathbf{x}$ e denota-se por $alr(\mathbf{x})$ à transformação $alr: S^D \to \mathbb{R}^{D-1}$, definida por

$$\mathbf{y} = alr(\mathbf{x}) = ln\left(\frac{x_1}{x_D}, \frac{x_2}{x_D}, \dots, \frac{x_{D-1}}{x_D}\right) \tag{2.10}$$

$$= (y_1, y_2, \dots, y_{D-1}),$$

em que $y_i = ln(x_i/x_D), i = 1, \dots, D-1$.

■

Além de x_D, qualquer outra parte da composição poderia ser escolhida como referência para figurar no denominador, conduzindo a diferentes transformações alr. Quando omisso, assume-se x_D no denominador.

Seja $\mathbf{y}$ as coordenadas alr-transformadas de uma composição de D partes $\mathbf{x} \in S^D$. A partir de $\mathbf{y}$ podemos obter a composição original $\mathbf{x}$, através da inversa da transformação alr, denotada por $alr^{-1}: \mathbb{R}^{D-1} \to S^D$, e definida por

$$\mathbf{x} = alr^{-1}(\mathbf{y}) = C(exp(y_1), exp(y_2), \dots, exp(y_{D-1}), 1),$$

onde $C(\cdot)$ denota a operação de fecho.

A transformação alr permite reduzir a perturbação e a potenciação a operações comuns de adição e multiplicação no espaço $\mathbb{R}^{D-1}$, pois, temos que

$$alr\big((\alpha \otimes \mathbf{x}) \oplus (\beta \otimes \mathbf{y})\big)$$

$$= alr\left((x_1^\alpha, x_2^\alpha, \dots, x_D^\alpha) \oplus (y_1^\beta, y_2^\beta, \dots, y_D^\beta)\right)$$

$$= alr\left(x_1^\alpha y_1^\beta, x_2^\alpha y_1^\beta, \dots, x_D^\alpha y_D^\beta\right)$$

$$= ln\left(\frac{x_1^\alpha y_1^\beta}{x_D^\alpha y_D^\beta}, \frac{x_2^\alpha y_1^\beta}{x_D^\alpha y_D^\beta}, \dots, \frac{x_{D-1}^\alpha y_{D-1}^\beta}{x_D^\alpha y_D^\beta}\right)$$

$$= \left(\alpha\, ln\left(\frac{x_1}{x_D}\right) + \beta ln\left(\frac{y_1}{y_D}\right), \alpha\, ln\left(\frac{x_2}{x_D}\right) + \beta ln\left(\frac{y_2}{y_D}\right), \dots, \alpha\, ln\left(\frac{x_{D-1}}{x_D}\right) + \beta ln\left(\frac{y_{D-1}}{y_D}\right)\right)$$

$$= \alpha \cdot alr(\mathbf{x}) + \beta alr(\mathbf{y}),$$

para quaisquer composições $\mathbf{x}$ e $\mathbf{y}$, e quaisquer constantes reais α e β.

Uma vez que a escolha de diferentes partes de referência para figurar no denominador resulta em diferentes transformações alr para uma mesma composição $\mathbf{x}$, isto significa que esta transformação não satisfaz o princípio de invariância de permutação e, portanto, a análise de dados composicionais através deste tipo de transformação pode resultar em conclusões pouco fidedignas (Pawlowsky-Glahn *et al.*, 2011).

2.3.3. Transformação clr

Para evitar problemas relacionados com a utilização de transformações alr, Aitchison (1986) introduziu a transformação de log-razões centradas, onde se representa uma composição de D partes através de D coordenadas clr, definida conforme se segue:

Definição 2.11 (Transformação clr)

Seja $\mathbf{x}$ uma composição de D partes no simplex S^D. Chama-se transformação log-razões centradas de $\mathbf{x}$, e denota-se por $clr(\mathbf{x})$, à transformação $clr: S^D \to U^D$ definida por

$$\mathbf{z} = clr(\mathbf{x}) = ln\left(\frac{x_1}{g(\mathbf{x})}, \frac{x_2}{g(\mathbf{x})}, \dots, \frac{x_D}{g(\mathbf{x})}\right), \tag{2.11}$$

em que $g(\mathbf{x}) = \left(\prod_{i=1}^{D} x_i\right)^{1/D}$ é a média geométrica de $\mathbf{x}$ e $U^D = \{(u_1, u_2, \dots, u_D) \in \mathbb{R}^D : u_1 + u_2 + \cdots + u_D = 0\}$ é um hiperplano de $\mathbb{R}^D$ (Pawlowsky-Glahn *et al.*, 2015).

■

A partir de $clr(\mathbf{x})$, a composição $\mathbf{x}$ pode ser recuperada pela inversa da transformação clr, denotada por $clr^{-1}: U^D \to S^D$, e definida por

$$\mathbf{x} = clr^{-1}(\mathbf{z}) = C(exp(z_1), exp(z_2), \dots, exp(z_D)), \tag{2.12}$$

De forma semelhante com o que acontece para a transformação alr, a perturbação e a potenciação em S^D correspondem à soma e ao produto no espaço real $\mathbb{R}^D$, ou seja, temos que

$$\begin{aligned}
clr(\alpha \otimes \mathbf{x} \oplus \beta \otimes \mathbf{y}) &= clr\left((x_1^\alpha, x_2^\alpha, \dots, x_D^\alpha) \oplus (y_1^\beta, y_2^\beta, \dots, y_D^\beta)\right) \\
&= clr\left(x_1^\alpha y_1^\beta, x_2^\alpha y_2^\beta, \dots, x_D^\alpha y_D^\beta\right) \\
&= ln\left(\frac{x_1^\alpha y_1^\beta}{g\left(x_1^\alpha y_1^\beta, \dots, x_D^\alpha y_D^\beta\right)}, \dots, \frac{x_D^\alpha y_D^\beta}{g\left(x_1^\alpha y_1^\beta, \dots, x_D^\alpha y_D^\beta\right)}\right) \\
&= ln\left(\frac{x_1^\alpha y_1^\beta}{\left(x_1^\alpha y_1^\beta \dots x_D^\alpha y_D^\beta\right)^{1/D}}, \dots, \frac{x_D^\alpha y_D^\beta}{\left(x_1^\alpha y_1^\beta \dots x_D^\alpha y_D^\beta\right)^{1/D}}\right) \\
&= ln\left(\frac{x_1^\alpha y_1^\beta}{(x_1^\alpha \dots x_D^\alpha)^{1/D}\left(y_1^\beta \dots y_D^\beta\right)^{1/D}}, \dots, \frac{x_D^\alpha y_D^\beta}{(x_1^\alpha \dots x_D^\alpha)^{1/D}\left(y_1^\beta \dots y_D^\beta\right)^{1/D}}\right)
\end{aligned}$$

$$= ln\left(\frac{x_1^{\alpha}y_1^{\beta}}{(x_1 \dots x_D)^{\alpha/D}(y_1 \dots y_D)^{\beta/D}}, \dots, \frac{x_D^{\alpha}y_D^{\beta}}{(x_1 \dots x_D)^{\alpha/D}(y_1 \dots y_D)^{\beta/D}}\right)$$

$$= ln\left(\left(\frac{x_1}{g(\mathbf{x})}\right)^{\alpha}\left(\frac{y_1}{g(\mathbf{y})}\right)^{\beta}, \dots, \left(\frac{x_D}{g(\mathbf{x})}\right)^{\alpha}\left(\frac{y_D}{g(\mathbf{y})}\right)^{\beta}\right)$$

$$= \left(\alpha ln\left(\frac{x_1}{g(\mathbf{x})}\right) + \beta ln\left(\frac{y_1}{g(\mathbf{y})}\right), \dots, \alpha ln\left(\frac{x_D}{g(\mathbf{x})}\right) + \beta ln\left(\frac{y_D}{g(\mathbf{y})}\right)\right)$$

$$= \alpha \cdot clr(\mathbf{x}) + \beta \cdot clr(\mathbf{y}).$$

Considerando que na transformação clr o denominador é a média geométrica das partes, então a análise de dados composicionais em coordenadas clr-transformadas satisfaz o princípio de invariância sob permutação. No entanto, visto que a média geométrica de uma composição completa não é necessariamente igual à média geométrica de uma de suas subcomposições, então não há garantia de coerência subcomposicional, o que pode resultar em correlações espúrias. Na verdade, as correlações entre componentes clr −transformadas não devem ser interpretadas como correlações entre variáveis originais. A desvantagem da transformação clr é que ela usa D coordenadas, para representar uma composição que tem apenas $D - 1$ componentes livres, que corresponde à dimensão de S^D.

Um aspecto muito importante sobre a representação de composições em coordenadas clr-transformadas é que ela pode ser usada para definir uma estrutura métrica no simplex. O produto interno, a norma e a distancia de Aitchison para composições em S^D são dadas, respetivamente, por

$$\langle \mathbf{x}, \mathbf{y} \rangle_a = \langle clr(\mathbf{x}), clr(\mathbf{y}) \rangle; \tag{2.13}$$

$$\|\mathbf{x}\|_a^2 = \|clr(\mathbf{x})\|^2, \quad d_a(\mathbf{x}, \mathbf{y}) = d\big(clr(\mathbf{x}), clr(\mathbf{y})\big) \tag{2.14}$$

onde $\langle \cdot, \cdot \rangle$, $\| \cdot \|$ *e* $d(\cdot, \cdot)$, denotam, respetivamente, o produto interno euclidiano, a norma e a distancia em R^D. Por exemplo, a distância de Aitchison entre duas composições $\mathbf{x}$ e $\mathbf{y}$ pertencentes a S^D é dada por

$$d_a(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{i=1}^{D}\big(clr_i(\mathrm{x}) - clr_i(\mathrm{y})\big)^2}$$

$$= \sqrt{\sum_{i=1}^{D}\left[ln\left(\frac{x_i}{y_i}\right) - ln\left(\frac{g(\mathbf{y})}{g(\mathbf{x})}\right)\right]^2} \tag{2.15}$$

Exemplo 2.5. Produto interno, distância e norma de Aitchison

Consideremos duas composições $\mathbf{x}, \mathbf{y} \in S^3$, tais que $\mathbf{x} = (0.25, 0.50, 0.25)$ e $\mathbf{y} = (0.50, 0.25, 0.25)$. Temos que:

(a) Coordenadas clr-transformadas das composições $\mathbf{x}$ e $\mathbf{y}$:

- $clr(\mathbf{x}) = ln\big[(0.25, 0.50, 0.25)/(0.25 \times 0.50 \times 0.25)^{1/3}\big]$

$$= ln\left(\frac{0.25}{0.315}, \frac{0.50}{0.315}, \frac{0.25}{0.315}\right) = ln(0.79, 1.59, 0.79)$$

- $clr(\mathbf{y}) = ln(1.59, 0.79, 0.79)$

(b) Produto interno de Aitchison entre $\mathbf{x}$ e $\mathbf{y}$:

$$\langle \mathbf{x}, \mathbf{y} \rangle_a = \langle ln(0.79, 1.59, 0.79), ln(1.59, 0.79, 0.79) \rangle$$
$$= ln(0.79) \times \ln(1.59) + \ln(1.59) \times \ln(0.79) + \ln(0.79) \times \ln(0.79) = -0{,}16$$

(c) Normas de Aitchison das composições $\mathbf{x}$ e $\mathbf{y}$:

- $\|\mathbf{x}\|_a = \sqrt{\langle clr(\mathbf{x}), clr(\mathbf{x}) \rangle}$
$$= \sqrt{\langle ln(0.79, 1.59, 0.79), \ln(0.79, 1.59, 0.79) \rangle} = \sqrt{0.33} = 0.57$$
- $\|\mathbf{y}\|_a = \sqrt{\langle clr(\mathbf{y}), clr(\mathbf{y}) \rangle} = 0.57$

(d) Distância de Aitchison entre as composições $\mathbf{x}$ e $\mathbf{y}$:

$$d_a(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{i=1}^{3} (clr_i(\mathbf{x}) - clr_i(\mathbf{y}))^2}$$
$$= \sqrt{[ln(0.79/1.59)]^2 + [ln(1.59/0.79)]^2 + [ln(0.79/0.79)]^2} = 0.99$$

■

Tal como acontece na geometria euclidiana, a norma e o produto interno de Aitchison permite-nos determinar o ângulo α entre dois vetores composicionais $\mathbf{x}$ e $\mathbf{y}$, a partir da seguinte relação:

$$\cos \alpha = \frac{\langle \mathbf{x}, \mathbf{y} \rangle_a}{\|\mathbf{x}\|_a^2 \cdot \|\mathbf{y}\|_a^2}. \tag{2.16}$$

O produto interno de Aitchison, a norma e a distância respeitam os princípios de análise composicional e constituem ferramentas para a análise composicional sem inconsistências. Esses operadores, juntamente com a perturbação e potenciação confere ao simplex a estrutura de um espaço euclidiano de dimensão $D - 1$, chamado geometria de Aitchison, satisfazendo as seguintes propriedades (Barceló-Vidal *et al*, 2003; Pawlowsky-Glahn *et al*, 2011):

(a) $d_a(\mathbf{p} \oplus \mathbf{x}, \mathbf{p} \oplus \mathbf{y}) = d_a(\mathbf{x}, \mathbf{y})$; (Preservação de distância sob perturbação)
(b) $d_a(\lambda \otimes \mathbf{x}, \lambda \otimes \mathbf{y}) = |\lambda| d_a(\mathbf{x}, \mathbf{y})$; (Distância e potenciação)
(c) $\mathbf{y} = \mathbf{x_0} \oplus (\alpha \otimes \mathbf{x})$. (Reta composicional, com origem em $\mathbf{x}_0$ e direção $\mathbf{x}$)

2.3.4. Transformações ilr

Um passo importante para trabalhar com a geometria de Aitchison consiste na criação de uma base ortonormal e suas correspondentes coordenadas.

Definição 2.12 (Base ortonormal no simplex)

Seja S^D o simplex de D partes. O conjunto de vetores $\{\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_{D-1}\}$, com $\mathbf{e}_i \in S^D, i = 1, 2, \dots, D - 1$, é uma base ortonormal de S^D se:

i. $\langle \mathbf{e}_i, \mathbf{e}_j \rangle_a = 0$ para $i \neq j$;
ii. $\|\mathbf{e}_i\|_a = 1, i = 1, 2, \dots, D - 1$.

■

Podemos encontrar facilmente um exemplo de base ortonormal em espaços vetoriais reais. Por exemplo, uma base em $\mathbb{R}^3$ é dada pelos vetores $(1,0,0),(0,1,0)$ *e* $(0,0,1)$, que é designada base canónica devido sua simplicidade. Mas, em S^3 não é assim tão simples! O procedimento para estabelecer uma base ortonormal no simplex foi proposto pela primeira vez por **Egozcue** *et al* em 2003. Por exemplo, uma base ortonormal em S^3 é dada pelos vetores (Buccianti *et al.,* 2006, pág. 153)

$$\mathbf{e}_1 = C\left[exp\left(\frac{1}{\sqrt{2}},\frac{-1}{\sqrt{2}},0\right)\right] = C\left(e^{\frac{1}{\sqrt{2}}},e^{\frac{-1}{\sqrt{2}}},1\right),$$

$$\mathbf{e}_2 = C\left[exp\left(\frac{1}{\sqrt{6}},\frac{1}{\sqrt{6}},\frac{-2}{\sqrt{6}}\right)\right] = C\left(e^{\frac{1}{\sqrt{6}}},e^{\frac{1}{\sqrt{6}}},e^{\frac{-2}{\sqrt{6}}}\right). \quad (2.17)$$

Em (2.17) os vetores $\left(\frac{1}{\sqrt{2}},\frac{-1}{\sqrt{2}},0\right)$ e $\left(\frac{1}{\sqrt{6}},\frac{1}{\sqrt{6}},\frac{-2}{\sqrt{6}}\right)$ correspondem, respetivamente, às coordenadas clr-transformadas da base formada pelos vetores $\mathbf{e}_1$ e $\mathbf{e}_2$, pelo que estes são obtidos pela transformação inversa de suas coordenadas clr-transformadas, tal como definida em (2.12), e satisfaz as propriedades da Definição 2.13:

1. $\langle \mathbf{e}_1, \mathbf{e}_2 \rangle_a = 0$; de facto
   - $\langle \mathbf{e}_1, \mathbf{e}_2 \rangle = \langle \left(\frac{1}{\sqrt{2}},\frac{-1}{\sqrt{2}},0\right), \left(\frac{1}{\sqrt{6}},\frac{1}{\sqrt{6}},\frac{-2}{\sqrt{6}}\right) \rangle$
     $= \frac{1}{\sqrt{2}} \times \frac{1}{\sqrt{6}} + \frac{-1}{\sqrt{2}} \times \frac{1}{\sqrt{6}} + 0 \times \frac{-2}{\sqrt{6}} = 0$

2. $\|\mathbf{e}_i\|_a = 1, i = 1,2$; de facto
   - $\|\mathbf{e}_1\|_a = \sqrt{\left(\frac{1}{\sqrt{2}}\right)^2 + \left(\frac{-1}{\sqrt{2}}\right)^2 + 0^2} = 1,$
   - $\|\mathbf{e}_2\|_a = \sqrt{\left(\frac{1}{\sqrt{6}}\right)^2 + \left(\frac{1}{\sqrt{6}}\right)^2 + \left(\frac{-2}{\sqrt{6}}\right)^2} = 1.$

Definição 2.13 (Transformação ilr)

Seja $\mathbf{x} \in S^D$ uma composição de D partes e $\{\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_{D-1}\}, \mathbf{e}_i \in S^D$, uma base ortonormal de S^D. Chama-se transformação de log-razões isométricas de $\mathbf{x}$ em relação à base $\{\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_{D-1}\}$, e denota-se por $ilr(\mathbf{x})$, à transformação $ilr: S^D \to \mathbb{R}^{D-1}$ dada por

$$\mathbf{x}^* = ilr(\mathbf{x}) = (x_1^*, x_2^*, \dots, x_{D-1}^*),$$

em que $x_i^* = \langle \mathbf{x}, \mathbf{e}_i \rangle_a$, ou seja,

$$x_i^* = \langle clr(\mathbf{x}), clr(\mathbf{e}_i) \rangle, \qquad i = 1,2,\dots,D-1. \quad (2.18)$$

■

Cada uma das coordenadas ilr-transformadas de $\mathbf{x} \in S^D$é obtida, portanto, pela projeção dessa composição sobre cada um dos vetores de uma dada base ortonormal do simplex S^D. Dadas as coordenadas ilr-transformadas de $\mathbf{x}$ em relação a uma base $\{\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_{D-1}\}$, podemos recuperar a composição original pela inversa da transformação ilr dada por $ilr^{-1}: \mathbb{R}^{D-1} \to S^{D-1}$ e definida por

$$\mathbf{x} = ilr^{-1}(\mathbf{x}^*) = C\big(exp(\langle \mathbf{x}^*, \mathbf{e}_1 \rangle_a, \langle \mathbf{x}^*, \mathbf{e}_2 \rangle_a, \dots, \langle \mathbf{x}^*, \mathbf{e}_{D-1} \rangle_a)\big). \quad (2.19)$$

Uma vez que coordenadas ilr-transformadas são obtidas a partir de uma base ortonormal, tal garante que a correspondência entre o simplex S^D e o espaço euclidiano $\mathbb{R}^{D-1}$ é isométrica[5].

Tal como na representação de composições em coordenadas clr-transformadas, a transformação ilr também pode ser usada para definir uma estrutura métrica no simplex, mas com a particularidade de que o produto interno, norma e distância entre vetores em coordenadas ilr correspondem ao espaço real $\mathbb{R}^{D-1}$ que é isomorfo a S^D. Assim, temos:

$$ilr((\alpha \otimes \mathbf{x}) \oplus (\beta \otimes \mathbf{y})) = \alpha \cdot ilr(\mathbf{x}) + \beta \cdot ilr(\mathbf{y}), \tag{2.20}$$

$$\langle \mathbf{x}, \mathbf{y} \rangle_a = \langle ilr(\mathbf{x}), ilr(\mathbf{y}) \rangle, \tag{2.21}$$

$$\|\mathbf{x}\|_a = \|ilr(\mathbf{x})\|, \tag{2.22}$$

$$d_a(\mathbf{x}, \mathbf{y}) = d(ilr(\mathbf{x}), ilr(\mathbf{y})). \tag{2.23}$$

Esta correspondência da métrica no espaço S^D à custa da métrica no espaço euclidiano permite a aplicação de técnicas usuais de análise multivariada aos dados composicionais em termos de suas coordenadas ilr-transformadas.

2.3.5. Base ortonormal baseada na Partição Binária Sequencial

Existem infinitas bases ortonormais em S^D. Algumas bases ortogonais especiais podem ser obtidas através de uma técnica conhecida por Partição Binária Sequencial (PBS), proposta inicialmente por Egozcue *et al* (2003). Esta técnica é definida em termos de uma partição predefinida das partes da composição. Tal implica que uma base construída usado a PBS dependerá da escolha dessa partição. Na literatura especializada são referidas duas formas de escolher partições com interesse prático. A primeira escolha baseia-se no conceito de equilíbrio (*balances*) entre grupos de partes e foi sugerida por Egozcue *et al* (2005), onde cada um dos $D-1$ passos da partição dá origem a uma coordenada ilr-transformada. A segunda escolha foi sugerida mais recentemente por Filzmoser et al (2009) com o interesse de garantir a interpretabilidade das coordenadas transformadas, concretamente, garantir que cada coordenada ilr-transformada explique todas as log-razões de uma variável original. A seguir, consideraremos a técnica da PBS segundo cada uma daquelas duas escolhas.

PBS segundo Egozcue *et al* (2005)

O processo de PBS baseado no conceito de equilíbrio entre grupos pode ser descrito da seguinte forma: na primeira etapa, divide-se a composição em dois grupos de partes, sendo as partes de um grupo etiquetadas por $+1$ e do outro por -1. A seguir, na segunda etapa, seleciona-se um dos grupos obtidos, o qual será novamente divido em dois grupos, seguindo-se o procedimento de etiquetagem aplicado na etapa anterior, sendo que este procedimento se repete até que todos os grupos sejam formados apenas por uma parte. As partes que não estiverem envolvidas na partição, numa certa etapa, serão etiquetadas com 0 (zero). Cada etapa de PBS está associada a um vetor $\mathbf{e}_i$ de uma base ortonormal e, consequentemente, tendo em conta (2.18), a uma coordenada ilr- transformada. Se o processo de agrupamento de partes se basear na afinidade exibida pelas partes tendo em conta o contexto dos dados (por exemplo, maiores e menores elementos, alcalinos e não alcalinos, seres vivos e não vivos, animais e vegetais, contaminantes e não contaminantes, etc.), as coordenadas geradas

[5] As distâncias entre composições em coordenadas transformadas são iguais às distâncias entre composições em coordenadas originais.

podem ser interpretadas em termos de peso relativo das partes em cada um dos grupos formados (Buccianti *et al*, 2006).

Sem perda de generalidade, assumimos que na i-ésima etapa da PBS um grupo de $r+s$ partes é dividido em dois grupos, sendo um formado por r partes (etiquetadas com $+1$) e outro formado por s partes (etiquetadas com -1). Nestas condições, o vetor da base ortonormal associada à i-ésima etapa da PBS é dado pela expressão

$$\mathbf{e}_i = C[exp(\Psi_{i1}, \Psi_{i2}, \ldots., \Psi_{iD})], \tag{2.24}$$

em que Ψ_{ij} corresponde à j-ésima coordenada clr-transformada do vetor $\mathbf{e}_i, i = 1, 2, \ldots, D-1$, associado a i-ésima etapa da PBS, e é dada por

$$\Psi_{ij} = \begin{cases} \sqrt{\dfrac{s}{r(r+s)}} & \text{, se etiqueta} = +1 \\ -\sqrt{\dfrac{r}{s(r+s)}} & \text{, se etiqueta} = -1 \\ 0 & \text{, se etiqueta} = 0 \end{cases} \tag{2.25}$$

Exemplo 2.6. Construção de uma base ortonormal usando PBS

Consideremos uma composição de quatro partes $\mathbf{x} = (x_1, x_2, x_3, x_4)$, em que cada uma das partes representa, respetivamente, a frequência dos nucleótidos A, T, C e G, nas três posições dos codões de uma dada espécie.

Na Tabela 2.5 podemos ver a partição binária sequencial para uma composição de quatro partes $\mathbf{x} = (x_1, x_2, x_3, x_4)$, em que cada uma das partes representa, respetivamente, as frequências dos nucleótidos A, T, C e G, nas três posições dos codões de uma dada espécie. Podemos realizar uma PBS de $\mathbf{x}$ do seguinte modo (Tabela 2.5):

- Etapa 1: etiqueta positiva para o par (A, T) e etiqueta negativa para o par (C, G);
- Etapa 2: etiqueta positiva para A, negativa para T e zero para as restantes;
- Etapa 3: etiqueta positiva para C, negativa para G e zero para as restantes.

Tabela 2.5. Partição binária sequencial de uma composição de 4 partes

Etapa	x_1	x_2	x_3	x_4	r	s
1	+1	+1	-1	-1	2	2
2	+1	-1	0	0	1	1
3	0	0	+1	-1	1	1

Os valores de Ψ_{ij} correspondentes às coordenadas clr-transformadas de vetores da base ortonormal do simplex S^4 obtidas por esta partição estão na Tabela 2.6.

Tabela 2.6. Valores de Ψ_{ij} associados ao processo de PBS de uma composição de 4 partes apresentado na Tabela 2.5.

Etapa	Ψ_{i1}	Ψ_{i2}	Ψ_{i3}	Ψ_{i4}
1	$\frac{1}{2}$	$\frac{1}{2}$	$-\frac{1}{2}$	$-\frac{1}{2}$
2	$\frac{1}{\sqrt{2}}$	$-\frac{1}{\sqrt{2}}$	0	0
3	0	0	$\frac{1}{\sqrt{2}}$	$-\frac{1}{\sqrt{2}}$

Então, a base obtida por essa partição é dada pelos vetores

$$\mathbf{e}_1 = C[\exp(\tfrac{1}{2}, \tfrac{1}{2}, -\tfrac{1}{2}, -\tfrac{1}{2})],$$

$$\mathbf{e}_2 = C[\exp(\tfrac{1}{\sqrt{2}}, -\tfrac{1}{\sqrt{2}}, 0, 0)],$$

$$\mathrm{e}_3 = C\left[\exp\left(0, 0, \frac{1}{\sqrt{2}}, -\frac{1}{\sqrt{2}}\right)\right]$$

■

A matriz formada pelos coordenadas $clr(\mathbf{e}_i), i = 1, 2, 3$, na Tabela 2.6 é designada por matriz de contrastes *(Contrast matrix)* associada à base ortonormal $\{\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3\}$, que, de um modo geral, é definida do seguinte modo (Pawlowsky-Glahn *et al.,* 2015):

Definição 2.14 (Matriz de contrastes)

Seja $\{\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_{D-1}\}$ uma base ortonormal do simplex S^D. Uma matriz $\boldsymbol{\Psi}_{D-1\times D} = [\Psi_{ij}]$, tal que a i-ésima linha $\Psi_i = clr(\mathbf{e}_i), i = 1, 2, \dots, D-1$, é chamada de matriz de contrastes associada à base ortonormal $\{\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_{D-1}\}$.

■

Da Definição 2.13, sabemos que a matriz $\boldsymbol{\Psi}_{D-1\times D}$ satisfaz a condição

$$\boldsymbol{\Psi} \cdot \boldsymbol{\Psi}' = \boldsymbol{I}_{D-1}. \tag{2.26}$$

Dada uma base ortonormal $\{\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_{D-1}\}$ obtida pela PBS, cada uma das coordenadas ortogonais de $\mathbf{x} = (x_1, x_2, \dots, x_D)$ é obtida pela projeção de $\mathbf{x}$ sobre cada um dos vetores $\boldsymbol{e}_i$, e usando (2.18) teremos:

$$x_i^* = \langle clr(\mathbf{x}), clr(\mathbf{e}_i)\rangle = \langle ln\left(\frac{x_1}{g(\mathbf{x})}, \frac{x_2}{g(\mathbf{x})}, \dots, \frac{x_D}{g(\mathbf{x})}\right), (\Psi_{i1}, \Psi_{i2}, \dots., \Psi_{iD})\rangle$$

$$= \sum_{j=1}^{D} ln\left(\frac{x_j}{g(\mathbf{x})}\right) \times \Psi_{ij} \tag{2.27}$$

$$= ln\left(\frac{(\prod x_+)^{\sqrt{s/(r(r+s))}}}{(\prod x_-)^{\sqrt{r/(s(r+s))}}}\right). \tag{2.28}$$

De facto, usando (2.25) e denotando por $x_i^{(+)}, i = 1, 2, \dots, r$, as r partes etiquetadas com $+1$ e por $x_i^{(-)}, i = 1, 2, \dots, s$ as s partes etiquetadas com -1 na i-ésima etapa da PBS, podemos escrever o somatório (2.27) do seguinte modo:

$$= \sum_{i=1}^{r} ln\left(\frac{x_i^{(+)}}{g(\mathbf{x})}\right) \times \sqrt{\frac{s}{r(r+s)}} + \sum_{i=1}^{s} ln\left(\frac{x_i^{(-)}}{g(\mathbf{x})}\right) \times \left(-\sqrt{\frac{r}{s(r+s)}}\right)$$

$$= \sum_{i=1}^{r} ln\left(\frac{x_i^{(+)}}{g(\mathbf{x})}\right) \times \sqrt{\frac{s \cdot r}{r(r+s) \cdot r}} - \sum_{i=1}^{s} ln\left(\frac{x_i^{(-)}}{g(\mathbf{x})}\right) \times \sqrt{\frac{r \cdot s}{s(r+s) \cdot s}}$$

$$= \sum_{i=1}^{r} ln\left(\frac{x_i^{(+)}}{g(\mathbf{x})}\right) \times \sqrt{\frac{rs}{r+s}} \times \frac{1}{r} - \sum_{i=1}^{s} ln\left(\frac{x_i^{(-)}}{g(\mathbf{x})}\right) \times \sqrt{\frac{rs}{r+s}} \times \frac{1}{s}$$

$$= \sqrt{\frac{rs}{r+s}} \left(\sum_{i=1}^{r} ln\left(\frac{x_i^{(+)}}{g(\mathbf{x})}\right)^{\frac{1}{r}} - \sum_{i=1}^{s} ln\left(\frac{x_i^{(-)}}{g(\mathbf{x})}\right)^{\frac{1}{s}} \right)$$

$$= \sqrt{\frac{rs}{r+s}} \left(\sum_{i=1}^{r} \left[ln\left(x_i^{(+)}\right)^{\frac{1}{r}} - ln(g(\mathbf{x}))^{\frac{1}{r}} \right] - \sum_{i=1}^{s} \left[ln\left(x_i^{(-)}\right)^{\frac{1}{s}} - ln(g(\mathbf{x}))^{\frac{1}{s}} \right] \right)$$

$$= \sqrt{\frac{rs}{r+s}} \left(\sum_{i=1}^{r} ln\left(x_i^{(+)}\right)^{\frac{1}{r}} - \sum_{i=1}^{s} ln\left(x_i^{(-)}\right)^{\frac{1}{s}} - r \times ln\left(g(\mathbf{x})\right)^{\frac{1}{r}} + s \times ln\left(g(\mathbf{x})\right)^{\frac{1}{s}} \right)$$

$$= \sqrt{\frac{rs}{r+s}} \left(\sum_{i=1}^{r} ln\left(x_i^{(+)}\right)^{\frac{1}{r}} - \sum_{i=1}^{s} ln\left(x_i^{(-)}\right)^{\frac{1}{s}} \right)$$

$$= \sqrt{\frac{rs}{r+s}} \left(ln\left(\prod_{i=1}^{r} x_i^{(+)} \right)^{\frac{1}{r}} - ln\left(\prod_{i=1}^{s} x_i^{(-)} \right)^{\frac{1}{s}} \right)$$

$$= \sqrt{\frac{rs}{r+s}}\, ln\left(\frac{\left(\prod_{i=1}^{r} x_i^{(+)} \right)^{1/r}}{\left(\prod_{i=1}^{s} x_i^{(-)} \right)^{1/s}} \right)$$

$$= ln\left(\frac{\left(\prod_{i=1}^{r} x_i^{(+)} \right)^{\sqrt{s/(r(r+s))}}}{\left(\prod_{i=1}^{s} x_i^{(-)} \right)^{\sqrt{r/(s(r+s))}}} \right).$$

Cada uma das coordenadas ortogonais x_i^* obtidas pelo processo de PBS é também chamada de equilíbrio entre os grupos de partes formados na i-ésima etapa de PBS (Egozcue *et al*, 2005).

Para ilustrar o processo de determinação de coordenadas ortogonais a partir de PSB retomemos o *Exemplo 2.3*. Com os dados do cientista A, que analisou a composição completa, formada por animal, vegetal, mineral e água, e com os dados do cientista B, que analisou uma subcomposição formada por animal, vegetal e mineral, executou-se o procedimento PBS para construir coordenadas ortogonais para as duas composições (Tabela 2.7 – cientista A; Tabela 2.8 – cientista B).

Detalhando para a Tabela 2.7, em cada etapa da PBS as partes foram agrupadas da seguinte forma:

- Etapa 1: etiqueta positiva para seres vivos e etiqueta negativa para seres não vivos;
- Etapa 2: etiqueta positiva para animal, negativa para vegetal e zero para mineral e água;
- Etapa 3: etiqueta positiva para mineral, negativa para água e zero para os seres vivos.

Tabela 2.7. Expressões de coordenadas ortogonais para uma composição de 4 partes obtida por PBS

Etapa	x_1	x_2	x_3	x_4	r	s	coordenadas ortogonais
1	+1	+1	-1	-1	2	2	$x_1^* = ln\left(\frac{(x_1x_2)^{1/2}}{(x_3x_4)^{1/2}}\right) = \frac{1}{2}ln\left(\frac{x_1x_2}{x_3x_4}\right)$
2	+1	-1	0	0	1	1	$x_2^* = ln\left(\frac{x_1{}^{1/\sqrt{2}}}{x_2{}^{1/\sqrt{2}}}\right) = \frac{1}{\sqrt{2}}ln\left(\frac{x_1}{x_2}\right)$
3	0	0	+1	-1	1	1	$x_3^* = ln\left(\frac{x_3{}^{1/\sqrt{2}}}{x_4{}^{1/\sqrt{2}}}\right) = \frac{1}{\sqrt{2}}ln\left(\frac{x_3}{x_4}\right)$

Tabela 2.8. Expressões Coordenadas ortogonais para uma composição de 3 partes obtida por PBS

Etapa	x_1	x_2	x_3	r	s	coordenadas ortogonais
1	+1	+1	-1	2	1	$x_1^* = ln\left(\frac{(x_1x_2)^{1/\sqrt{6}}}{x_3^{\sqrt{2/3}}}\right)$
2	+1	-1	0	1	1	$x_2^* = ln\left(\frac{x_1{}^{1/\sqrt{2}}}{x_2{}^{1/\sqrt{2}}}\right)$

Com base nas expressões de coordenadas ortogonais construídas nas Tabelas 2.7 e 2.8 podemos reescrever os dados das amostras recolhidas pelos cientistas A e B agora em termos de suas coordenas ilr-transformadas (Tabela 2.9).

Tabela 2.9. Dados em coordenadas ortogonais registadas pelos cientistas A e B

Cientista A			Cientista B	
x_1^*	x_2^*	x_3^*	x_1^*	x_2^*
-0.5493061	-0.4901291	-1.266965	0.2829762	-0.4901291
-0.5493061	0.4901291	-1.266965	-0.2829762	0.4901291
0.4054651	0.000000	0.000000	0.9162257	0.00000

Das Definições 2.12 e 2.13 podemos observar que existe uma relação linear entre as coordenadas ilr −transformadas e clr −transformadas. Concretamente, dada uma matriz de contrastes $\boldsymbol{\Psi}_{D-1\times D}$, as coordenadas ilr-transformadas de $\mathbf{x}$, conforme definida em (2.18), podem ser escritas na forma matricial do seguinte modo:

$$\mathbf{x}^* = clr(\mathbf{x}) \cdot \boldsymbol{\Psi}^T, \tag{2.28}$$

De forma semelhante, dadas as coordenadas $\mathbf{x}^* = ilr(\mathbf{x})$ em relação a uma base ortonormal, cujas coordenadas clr-transformadas são as entradas da matriz $\boldsymbol{\Psi}_{D-1\times D}$, podemos recuperar as coordenadas clr-transformadas da composição original $\mathbf{x} \in S^D$ pela seguinte relação:

$$clr(\mathbf{x}) = \mathbf{x}^* \cdot \boldsymbol{\Psi}_{D-1\times D}. \tag{2.29}$$

Esta relação linear entre as transformações clr e ilr é muito importante, pois permite que os resultados de uma análise realizada com dados em coordenadas ilr-trasnformadas sejam facilmente convertidos para serem interpretados no espaço clr-transformado, sem perda de informação.

PBS segundo Filzmoser *et al* (2009)

Na equação 2.27 podemos observar as coordenadas de uma composição em relação à uma base ortonormal apresentam uma relação muito complexa com as variáveis originais. Assim, para efeito de análises realizadas em coordenadas ilr-transformadas, pode ser muito difícil interpretar os resultados em termos das variáveis originais. Por outro lado, quando a base ortonormal se baseia na PBS proposta por Egozcue *et al* (2005), a nossa capacidade de fazer a separação de partes de modo que seja interpretável depende do nosso conhecimento a priori sobre o problema em estudo. A PBS tende a ficar confusa para composições que envolvem muitas partes e/ou quando nenhuma informação a *priori* sobre o problema está disponível, o que pode condicionar a eficácia de análise com dados em coordenadas ilr-transformadas na prática (Hron, 2012). Com o objetivo de ultrapassar esse constrangimento, Filzmoser *et al* (2009) propõe uma escolha adequada de bases de modo que cada uma das coordenadas ortogonais explique todas as log-razões de uma variável original.

Dada uma composição $\mathbf{x} \in S^D$, denotemos $\mathbf{x}^{(l)} = (x_1^{(l)}, x_2^{(l)}, \dots, x_D^{(l)})$ uma permutação de $\mathbf{x}$. Se executarmos D permutações tal que na l-ésima permutação a parte $x_l, l = 1, 2, \dots, D$, de $\mathbf{x}$ ocupe a primeira posição, obtemos D vetores composicionais que contêm a mesma informação relativa de $\mathbf{x}$, definidos por $\mathbf{x}^{(l)} = (x_l, x_1, x_2, \dots, x_{l-1}, x_{l+1}, \dots, x_D)$. Por exemplo, para $D = 4$, permutando as partes de uma composição $\mathbf{x} \in S^4$, os quatro vetores seguintes correspondem às quatro permutações desejadas e contêm a mesma informação relativa de $\mathbf{x}$**:**

$$\mathbf{x}^{(1)} = \mathbf{x} = (x_1, x_2, x_3, x_4);$$

$$\mathbf{x}^{(2)} = (x_2, x_1, x_3, x_4);$$

$$\mathbf{x}^{(3)} = (x_3, x_1, x_2, x_4);$$

$$\mathbf{x}^{(4)} = (x_4, x_1, x_2, x_3).$$

Se considerarmos uma partição do vetor $\mathbf{x}^{(l)}$ de modo que na primeira partição tenhamos um grupo formado pela componente $x_1^{(l)}$ e outro grupo formado pelas partes $x_2^{(l)}, x_3^{(l)}, x_4^{(l)}, \dots, x_D^{(l)}$; na segunda partição separamos o grupo $\{x_2^{(l)}, x_3^{(l)}, x_4^{(l)}, \dots, x_D^{(l)}\}$ obtido na etapa anterior de modo que tenhamos um grupo formado pela parte $x_2^{(l)}$ e o outro formado pelas partes $x_3^{(l)}, x_4^{(l)}, \dots, x_D^{(l)}$; e procedendo deste

modo até à partição de ordem $D-1$, em que teremos um grupo formado pela parte $x_{D-1}^{(l)}$ e outro pela parte $x_D^{(l)}$.

Na Tabela 2.10 apresentamos a PBS para esta partição e a correspondente matriz de contraste contendo as coordenadas clr-transformadas dos vetores $\mathbf{e}_i$ associados a cada etapa $i = 1, 2, \dots, D-1$, da partição (Equação 2.24). De acordo com as expressões de coordenadas clr-transformadas dos vetores da base ortonormal obtidos pela PBS, o vetor $\mathbf{e}_i, i = 1, 2, \dots, D-1$, associado à i-ésima ordem desta partição é dado por

$$\mathbf{e}_i = C[exp(\Psi_i)],$$

em que

$$\Psi_i = clr(\mathbf{e}_i) = \left(\underbrace{0, 0, .., 0}_{i-1\ elementos}, \underbrace{\sqrt{\frac{D-i}{D-i+1}}}_{1\ elemento}, \underbrace{\frac{-1}{\sqrt{(D-i)(D-i+1)}}, \dots, \frac{-1}{\sqrt{(D-i)(D-i+1)}}}_{D-i\ elementos} \right). \quad (2.29)$$

E, cada uma das coordenadas ortogonais de $\mathbf{x}^{(l)}, l = 1, 2, \dots, D$, conforme definida em (2.26), em relação à base ortonormal da PBS apresentada na Tabela 2.10, é um vetor $\boldsymbol{z}^{(l)} = \left(z_1^{(l)}, z_2^{(l)}, \dots, z_{D-1}^{(l)}\right), l = 1, 2, \dots, D$, cujas coordenadas $z_i^{(l)}$ associadas ao vetor $\mathbf{e}_i, i = 1, 2, \dots, D-1$, são dadas por

$$z_i^{(l)} = \sqrt{\frac{D-i}{D-i+1}}\, ln\left(\frac{x_i^{(l)}}{\left(\prod_{j=i+1}^{D} x_j^{(l)}\right)^{\frac{1}{D-i}}} \right), \qquad i = 1, 2, \dots, D-1. \quad (2.30)$$

De acordo com (2.30), a primeira coordenada ilr −transformada $z_1^{(l)}$ contém toda informação relativa entre a parte x_l e as restantes partes da composição original. Deste modo, uma vez que $x_1^{(l)} = x_l$ então $x_1^{(l)}$ corresponderá à única posição importante visto que pode ser explicada por $z_1^{(l)}, l = 1, 2, \dots, D$. Deste modo, a interpretação de resultados da análise de dados composicionais em termos de coordenadas ortogonais pode ser feita em termos de coordenadas originais (Filzmoser *et al,* 2011; Hron, 2012).

Outra vantagem da escolha de bases usando a PBS segundo a metodologia proposta por Filzmoser (Tabela 2.10) prende-se com a relação existente entre as coordenadas ortonormais $z_1^{(l)}, l = 1, 2, \dots, D$, e as coordenadas clr-transformadas de uma dada composição $\mathbf{x} \in S^D$. Por exemplo, se considerarmos a permutação $\mathbf{x}^{(l)} = (x_l, x_1, \dots, x_{l-1}, x_{l+1}, \dots, x_D)$, podemos escrever a l −ésima coordenada clr-transformada de $\mathbf{x}$ do seguinte modo:

$$y_l = ln \frac{x_1^{(l)}}{\left(\prod_{i=1}^{D} x_i\right)^{\frac{1}{D}}} \quad (2.31)$$

$$= \sqrt{\frac{D-1}{D}}\, z_1^{(l)}, l = 1, 2, \dots, D. \quad (2.32)$$

Tabela 2.10. PBS para construção de uma base ortonormal onde a primeira partição confronta a parte $x_1^{(l)}$ com as restantes partes da composição.

Resultados de cada etapa da PBS

Etapa da PBS	$x_1^{(l)}$	$x_2^{(l)}$	$x_3^{(l)}$	$x_4^{(l)}$	$\cdots$	$x_{D-2}^{(l)}$	$x_{D-1}^{(l)}$	$x_D^{(l)}$	r	s
1	$+1$	-1	-1	-1	$\cdots$	-1	-1	-1	1	$D-1$
2	0	$+1$	-1	-1	$\cdots$	-1	-1	-1	1	$D-2$
3	0	0	$+1$	-1	$\cdots$	-1	-1	-1	1	$D-3$
4	0	0	0	$+1$	$\cdots$	-1	-1	-1	1	$D-4$
$\vdots$			$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$			$\vdots$
$D-2$	0	0	0	0	$\cdots$	$+1$	-1	-1	1	2
$D-1$	0	0	0	0	$\cdots$	0	$+1$	-1	1	1

Matriz de contrastes

Etapa da PBS	Ψ_{i1}	Ψ_{i2}	Ψ_{i3}	$\cdots$	Ψ_{iD-1}	Ψ_{iD}	Ψ_i
1	$\sqrt{\frac{D-1}{1\cdot(1+D-1)}}$	$\frac{-1}{\sqrt{(D-1)\cdot(1+D-1)}}$	$\frac{-1}{\sqrt{(D-1)\cdot(1+D-1)}}$	$\cdots$	$\frac{-1}{\sqrt{(D-1)\cdot(1+D-1)}}$	$\frac{-1}{\sqrt{(D-1)\cdot(1+D-1)}}$	$\Psi_1^{(l)}$
2	0	$\sqrt{\frac{D-2}{1\cdot(1+D-2)}}$	$\frac{-1}{\sqrt{(D-2)\cdot(1+D-2)}}$	$\cdots$	$\frac{-1}{\sqrt{(D-1)\cdot(1+D-1)}}$	$\frac{-1}{\sqrt{(D-2)\cdot(1+D-2)}}$	$\Psi_2^{(l)}$
3	0	0	$\sqrt{\frac{D-2}{1\cdot(1+D-2)}}$	$\cdots$	$\frac{-1}{\sqrt{(D-1)\cdot(1+D-1)}}$	$\frac{-1}{\sqrt{(D-2)\cdot(1+D-2)}}$	$\Psi_3^{(l)}$
$\vdots$			$\vdots$	$\ddots$	$\vdots$		$\vdots$
$D-1$	0	0	0	$\cdots$	$\sqrt{\frac{1}{1\cdot(1+1)}}$	$\frac{-1}{\sqrt{1\cdot(1+1)}}$	$\Psi_{D-1}^{(l)}$

De facto, rearranjando os fatores em (2.31) podemos escrever y_l de forma equivalente conforme se segue:

$$y_l = ln\frac{x_1^{(l)}}{\left(\prod_{i=2}^{D} x_i^{(l)}\right)^{\frac{1}{D}} \cdot \left(x_1^{(l)}\right)^{\frac{1}{D}}} = ln\frac{\left(x_1^{(l)}\right)^{1-\frac{1}{D}}}{\left(\prod_{i=2}^{D} x_i^{(l)}\right)^{\frac{1}{D}}}$$

$$= ln\frac{\left(x_1^{(l)}\right)^{1-\frac{1}{D}}}{\left(\prod_{i=2}^{D} x_i^{(l)}\right)^{\frac{1}{D}}} = ln\frac{\left(x_1^{(l)}\right)^{\frac{D-1}{D}}}{\left(\left(\prod_{i=2}^{D} x_i^{(l)}\right)^{\frac{1}{D-1}}\right)^{\frac{D-1}{D}}}$$

$$= ln\left(\frac{x_1^{(l)}}{\left(\prod_{i=2}^{D} x_i^{(l)}\right)^{\frac{1}{D-1}}}\right)^{\frac{D-1}{D}} = \frac{D-1}{D} ln \frac{x_1^{(l)}}{\left(\prod_{i=2}^{D} x_i^{(l)}\right)^{\frac{1}{D-1}}}$$

$$= \sqrt{\frac{D-1}{D}} z_1^{(l)}, l = 1, 2, \dots, D.$$

Assim, concluímos que y_l é proporcional a $z_1^{(l)}$ e, portanto, cada uma das coordenadas ilr-transformadas $z_1^{(l)}, l = 1, 2, \dots, D$, podem ser interpretadas da mesma forma que as correspondentes coordenadas clr-transformadas $y_l, l = 1, 2, \dots, D$, visto que ambas explicam as log-razões correspondentes à l −ésima parte da composição (Kynclová *et al*, 2015).

Em síntese, as transformações log-contrastes, alr, clr e ilr, devem ser tidas em conta na análise de dados composicionais. De uma forma geral, a filosofia de análise de log-contrastes pode ser resumida em cinco passos que são (Aitchison, 2005):

1. Formulação do problema em termos de componentes da composição;
2. Tradução desta formulação em termos de vetores de log-contrastes da composição;
3. Transformação dos dados composicionais em vetores de log-contrastes;
4. Análise dos dados expressos em log-contrastes por uma técnica usual apropriada de análise multivariada;
5. Interpretação dos resultados obtidos no passo 4 em termos de log-contrastes de composições e em termos das variáveis originais.

Em relação à escolha da transformação a usar, apresentamos um quadro-resumo das transformações log-contrastes acima apresentadas, realçando vantagens e desvantagens de cada uma. A escolha do log-contraste a utilizar em cada situação dependerá dos objetivos do analista.

Quadro resumo das transformação log-razões alr, clr e ilr

Transformações	Vantagens	Desvantagens
alr *(Additive logratio)*: Transformação baseada no logaritmo de razões, com numa única variável de referência no denominador. $alr(\mathbf{x}) = ln\left(\frac{x_1}{x_D}, \dots, \frac{x_{D-1}}{x_D}\right)$	Reduz operações de perturbação e potenciação no simplex S^D a correspondentes operações de adição e multiplicação por um escalar no espaço euclidiano $\mathbb{R}^{D-1}$.	Não é isométrica e, portanto, não satisfaz o princípio de invariância de permutação, nem permite o cálculo de distâncias e produto interno no espaço euclidiano em $\mathbb{R}^{D-1}$.
clr *(Centered logratio)*: transformação isométrica baseada no logaritmo de razões em relação à média geométrica das variáveis. $clr(\mathbf{x}) = ln\left(\frac{x_1}{g(\mathbf{x})}, \dots, \frac{x_D}{g(\mathbf{x})}\right)$	Evita a escolha de uma proporção variável como acontece no alr, e simplifica a interpretação das variáveis transformadas, visto que permite analisar em termos das variáveis originais.	Os dados transformações apresentam incoerência subcomposicional, e resulta em matriz de dados singulares, o que inviabiliza a aplicação de técnicas robustas para dados nessas coordenadas.
ilr *(isometric logratio)*: Transformação isométrica baseada na escolha de uma base ortonormal $\{\mathbf{e}_1, \dots, \mathbf{e}_{D-1}\}$ no hiperplano formado por coordenadas clr-transformadas de $\mathbf{e}_1, i = 1, 2, \dots, D-1$. $ilr(\mathbf{x}) = \langle \mathbf{x}, \mathbf{e}_i \rangle_a$	Preserva todas as propriedades vantajosas da transformação clr e satisfaz todos os princípios de análise composicional.	Correlações calculadas com base na transformação ilr não podem ser interpretadas de acordo com as variáveis originais, visto que as variáveis ilr-transformadas estão relacionadas com variáveis originais através de funções não lineares.

CAPÍTULO 3

GRUPOS DE PARTES DE DADOS COMPOSICIONAIS

3.1. Introdução

Como vimos, os dados composicionais são de natureza multivariada. No entanto, a necessidade de se interpretar dados composicionais em termos de razões entre as partes ou log-razões das partes torna a análise muito mais complicada em relação à interpretação de dados em termos de variação absoluta, como é usual na Análise Estatística Multivariada. Ao lidar com esse tipo de dados, muitas vezes, por questões práticas, temos a necessidade de reduzir a dimensão dos dados, sem grande perda de informação, de modo a obter uma nova composição com a qual seja mais fácil trabalhar. Assim, Aitchison (1986) introduziu o conceito de fusão *(amalgamation)* de dados composicionais, com o objetivo de reduzir a dimensão dos dados ou para evitar a existência de componentes com valor zero. No entanto, após a constatação de que a operação de fusão introduzida por Aitchison (1986) não é compatível com a geometria de Aitchison, Pawlowsky-Glahn *et al* (2005) introduziu o conceito de equilíbrios, que além de servir para reduzir a dimensão dos dados, visa facilitar a interpretação dos resultados da análise com base em grupos formados pelas componentes de uma composição (Mateu-Figueras *et al*, 2008).

3.2. Fusão

Definição 3.1 (Fusão)

Seja $\mathbf{x} \in S^D$ uma composição de D partes. Chamamos fusão (*amalgamation*) de $\mathbf{x}$ a uma composição de d partes, obtida pela separação das partes de $\mathbf{x}$ em $d(\leq D)$ subconjuntos mutuamente exclusivos e exaustivos, somando-se componentes de cada subconjunto.

■

Exemplo 3.1. Fusão de uma composição (Espaço dos codões)

Consideremos o espaço dos codões onde cada composição é um vetor de 12 partes $\mathbf{x} = (x_1, x_2, \dots, x_{12})$, conforme definido no Capítulo 1. Se pretendemos estudar a composição dos quatro nucleótidos (A, C, G, T) nos codões das 31 espécies em estudo (Tabela 1.1) independentemente da posição em que ocorrem, podemos analisar a fusão $\boldsymbol{a} = (a_1, a_2, a_3, a_4)$, em que cada uma das componentes de $\boldsymbol{a}$ corresponde à soma das frequências de cada uma das quatro bases nas três posições dos codões, ou seja, $a_1 = x_1 + x_5 + x_9, a_2 = x_2 + x_6 + x_{10}$, $a_3 = x_3 + x_7 + x_{11}$ e $a_4 = x_4 + x_8 + x_{12}$. Deste modo, a composição original de 12 componentes fica reduzida a uma composição de apenas 4 componentes.

■

Na forma matricial, a fusão $\boldsymbol{a}$ será dada por

$$\boldsymbol{a} = \boldsymbol{A}\mathbf{x},$$

onde

$$\boldsymbol{A} = \begin{bmatrix} 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \end{bmatrix}$$

é chamada de matriz de fusão *(amalgamation matrix)*.

Definição 3.2 (Matriz de fusão)

Seja $\mathbf{x}$ uma composição de D partes. Qualquer matriz de ordem $d \times D$ $(d \leq D)$ com D elementos iguais a 1 (um), estando apenas um em cada coluna e no mínimo um em cada linha, e com os restantes $(d-1) \times D$ elementos iguais a 0 (zero), chama-se matriz de fusão relativamente a uma fusão de d partes extraídas da composição $\mathbf{x}$.

■

A fusão apresenta as seguintes propriedades (Aitchison, 1986):

1. Se uma composição de D partes $\mathbf{x}$ for multiplicada por uma matriz de fusão $\boldsymbol{A}_{\boldsymbol{d}\times\boldsymbol{D}}$, a fusão resultante $\boldsymbol{a} = \boldsymbol{A}\mathbf{x}$ é uma composição de d partes, ou seja, a matriz de fusão é uma transformação $A: S^D \rightarrow S^d$;
2. Toda matriz de permutação $D \times D$, incluindo a matriz identidade $\boldsymbol{I}_D$, é uma matriz de fusão.

Uma fusão deve ser aplicada apenas na fase de definição do problema em estudo, onde escolhemos as partes que serão consideradas e as unidades em que serão representadas. Assim, uma vez escolhidas as partes, as unidades e os objetivos da análise, já não se deve fundir mais variáveis. Por isso, a fusão deve ser feita de tal forma que facilite a interpretação dos resultados da análise, visto que, futuramente, não será possível alterá-la (Boogaart, K. G. *et al*, 2013).

Por vezes, podemos estar interessados em considerar não só a fusão de dados, mas também as subcomposições formadas por cada grupo de partes fundidas. Para isso, Aitchison (1986) introduziu o conceito de partição de uma composição, que pode ser definida do seguinte modo:

Definição 3.2 (Partição)

Seja $\mathbf{x} \in S^D$ uma composição de D partes da qual extraímos uma fusão de d $(\leq D)$ partes. Chamamos partição de ordem[6] $c = d - 1$ de $\mathbf{x}$ à fusão conjuntamente com as subcomposições associadas a cada um dos subconjuntos das partes de $\mathbf{x}$ que definiram a fusão.

■

Exemplo 3.2. Partição de uma composição (Espaço dos codões)

Retomemos o *Exemplo 3.1,* onde definimos uma fusão para uma composição do espaço dos codões, dada por $\boldsymbol{a} = (a_1, a_2, a_3, a_4)$, em que cada uma das componentes de $\boldsymbol{a}$ corresponde à soma das frequências de cada uma das quatro bases nas três posições do codão, ou seja, $a_1 = x_1 + x_5 + x_9, a_2 = x_2 + x_6 + x_{10}$, $a_3 = x_3 + x_7 + x_{11}$ e $a_4 = x_4 + x_8 + x_{12}$. No espaço dos codões podemos também estar interessados na proporção do nucleótido A nas três posições dos codões das espécies observadas, o que corresponde à subcomposição $\boldsymbol{s}_1 = C(x_1, x_5, x_9)$. De forma similar, podemos considerar as subcomposições formadas apenas pelo nucleótido C, G ou T, respetivamente, $\boldsymbol{s}_2 = C(x_2, x_6, x_{10})$, $\boldsymbol{s}_3 = C(x_3, x_7, x_{11})$ e $\boldsymbol{s}_4 = C(x_4, x_8, x_{12})$.

[6] A ordem de uma partição $\mathbf{x} \in S^D$ corresponde ao número de barras verticais necessárias para separar $\mathbf{x}$ em d partes (Aitchison, 1986).

Esta consideração simultânea da fusão $\boldsymbol{a} \in S^4$ e das subcomposições $\boldsymbol{s}_i, i = 1,2,3,4$, constitui uma partição de quatro partes da composição $\mathbf{x} = (x_1, x_2, \dots, x_{12})$.

■

Quando o nosso interesse consiste no estudo da relação entre grupos mutuamente exclusivos e exaustivos da composição tal como $(x_1 + x_2 + \cdots + x_r)/(x_{r+1}, x_{r+2} + \cdots + x_D)$, para algum r, a fusão de partes de uma composição constitui uma operação útil para a redução da dimensão de dados. No entanto, após a fusão dos dados, já não nos será possível analisar a relação entre as partes de cada grupo, pelo que a interpretação de resultados de uma análise feita sobre dados fundidos, em termos de variáveis originais, pode ser difícil.

Outro fator que merece o nosso cuidado ao usar a fusão como técnica de redução da dimensão de dados composicionais está no facto de que a fusão não é compatível com a geometria de Aitchison no simplex, conforme ilustraremos no exemplo que se segue.

Exemplo 3.3. A fusão de dados não preserva a distância de Aitchison sob a perturbação (Egozcue *et al* (2005), pág. 799):

Consideremos duas composições de três partes, sendo $\boldsymbol{a} = (0.1, 0.8, 0.1)$ e $\boldsymbol{b} = (0.3, 0.6, 0.1)$. Pretendemos comparar as distâncias de Aitchison entre essas duas composições após a perturbação por uma terceira composição $\boldsymbol{c} = (0.2, 0.7, 0.1)$ para dados fundidos e não fundidos.

Tabela 3.1 Efeito da perturbação das composições $\boldsymbol{a}$ e $\boldsymbol{b}$ pela composição $\boldsymbol{c} = (0.2, 0.7, 0.1)$ na distância de Aitchison, d_a, antes e depois da fusão.

		Não fundidos				Fundidos		
	Comp	x_1	x_2	x_3	$d_a(\boldsymbol{a}, \boldsymbol{b})$	$x_1 + x_2$	x_3	$d_a(\boldsymbol{a}, \boldsymbol{b})$
Não perturbados	$\boldsymbol{a}$	0.1	0.8	0.1	**1.035**	0.9	0.1	**0.00**
	$\boldsymbol{b}$	0.3	0.6	0.1		0.9	0.1	
Perturbados	$\boldsymbol{a}$	0.034	0.949	0.017	**1.035**	0.983	0.017	**0.134**
	$\boldsymbol{b}$	0.123	0.857	0.020		0.980	0.020	

Na Tabela 3.1 podemos observar que para dados fundidos (obtidos pela soma das partes 1 e 2), a distância de Aitchison em S^2 altera-se após a perturbação das composições, enquanto que a distância entre dados não fundidos mantém-se inalterada após a perturbação. Portanto, uma análise sobre dados fundidos pode conduzir a conclusões completamente diferentes daqueles obtidos com base na análise dos dados originais.

■

Esta incompatibilidade da perturbação com a geometria de Aitchison levou Egozcue *et al* (2005) a introduzir o conceito de equilíbrio entre grupos, que além de servir para reduzir a dimensão dos dados, visa facilitar a interpretação de resultados de análises tomando grupos formados pelas componentes de uma composição (Mateu-Figueras *et al*, 2008).

3.3. Equilíbrio

Já referimos que a análise de dados composicionais baseia-se nas log-razões entre as partes da composição, porque a única informação relevante para a análise composicional é a proporção das partes. Assim, muitas vezes, temos a necessidade de interpretar os resultados em termos de log-razões (ou razões) entre partes. Com o objetivo de facilitar a análise, torna-se conveniente que os dados sejam organizados de tal modo que possam ser agrupados em dois ou mais subconjuntos, que sejam interpretáveis de alguma forma. Ao analisar uma composição, podemos estar interessados em estudar as caraterísticas das composições da amostra de duas formas:

(a) A relação ou equilíbrio entre esses grupos de partes, conhecida como análise inter-grupos;
(b) O comportamento das partes em cada grupo, conhecida como análise intra-grupo.

Os grupos de partes podem ser vistos tanto como uma subcomposição, quer como um grupo dentro da composição completa. No entanto, a análise subcomposicional destina-se a estudar partes dentro de um mesmo grupo, sem se preocupar com as relações de um dado grupo com os restantes grupos. Uma análise composicional realizada nesta perspetiva corresponde à análise intra-grupo, e será considerada nas subseções 4.2.2 e 4.2.5, onde abordaremos diagramas ternários e biplots como técnicas de visualização da estrutura dos dados composicionais.

Por seu turno, a análise inter-grupos baseia-se, geralmente, nos conceito de fusão e de equilíbrio[7] (*balances*) entre grupos. O conceito de equilíbrio surge no processo de PBS de uma dada composição. Embora cada uma das coordenadas ortogonais x_i^*, obtida na i-ésima etapa da PBS corresponde ao equilíbrio entre os grupos de partes formados nesta etapa, aqui, denotaremos cada uma coordenadas por b_i, para realçar o facto de que o nosso interesse está na diferença relativa entre os grupos de partes e não nas coordenadas (ver a Subseção 2.3.5). Portanto, de acordo com (2.27), o equilíbrio entre os grupos de partes formados na i-ésima etapa da PBS de uma dada composição é dado por

$$b_i = \sqrt{\frac{rs}{r+s}}\, ln \frac{\left(\prod_{i=1}^{r} x_i^{(+)}\right)^{1/r}}{\left(\prod_{i=1}^{s} x_i^{(-)}\right)^{1/s}}. \tag{3.7}$$

Dada a sua forma de construção, a utilização de equilíbrios permite a comparação entre comportamento de dois grupos de partes e é compatível com a geometria de Aitchison (Egozcue *et al,* 2005).

De acordo com Pawlowsky-Glahn *et al* (2015), a interpretação dos equilíbrios pode ser feita com base em algumas das suas propriedades. Por exemplo, em (3.7) podemos observar a utilização de médias geométricas como representantes dos grupos no numerador e no denominador. Visto que as médias geométricas são valores centrais das partes de cada grupo, a razão entre as médias (geométricas) indica o peso relativo de cada grupo. Assim, por exemplo, um equilíbrio positivo significa que, em média (geométrica), o grupo de partes no numerador é dominante, pois tem maior peso na composição do que o grupo no denominador e o valor absoluto desse equilíbrio indica a diferença entre os grupos numa escala log-relativa.

[7] Optamos por traduzir o termo *balance* por equilíbrio. Na verdade, trata-se de uma medida que avalia a posição do fulcro de uma balança: quando equilibrada o fulcro está na posição zero.

Exemplo 3.5. Equilíbrios entre grupos

Para a interpretação de equilíbrios entre grupos, vamos considerar os dados da Tabela 2.1 referentes a uma amostra de composições de 4 partes, registada pelo cientista A referido no Exemplo 2.3. Nas Tabelas 2.5 e 2.6 já tínhamos aplicado PBS para separar a composição em grupos e determinado a matriz de contrastes. Aqui, na Tabela 3.2, apresentamos a PBS obtida e a expressão de equilíbrio obtida em cada etapa de PBS.

Na Tabela 3.2 podemos observar os grupos formados em cada etapa da PBS e as expressões que nos permitem calcular o equilíbrio entre os grupos formados nessas etapas. Devemos ter em atenção que os diferentes equilíbrios correspondem às coordenadas ilr-transformadas para os dados. E, neste caso, os valores calculados correspondem aos da Tabela 2.7, que voltamos a colocar aqui para facilitar a análise dos resultados.

Tabela 3.2. Equilíbrios entre grupos da composição de 4 partes referentes a amostras de solo registadas pelo cientista A do Exemplo 2.3.

Etapa	x_1	x_2	x_3	x_4	r	s	Equilíbrios entre grupos
1	+1	+1	-1	-1	2	2	$b_1 = \sqrt{\frac{2\times 2}{2+2}} \times ln\left[\frac{(x_1x_2)^{1/2}}{(x_3x_4)^{1/2}}\right] = \frac{1}{2} ln\left(\frac{x_1x_2}{x_3x_4}\right)$
2	+1	-1	0	0	1	1	$b_2 = \sqrt{\frac{1\times 1}{1+1}} \times ln\left[\frac{{x_1}^{1/1}}{{x_2}^{1/1}}\right] = \frac{1}{\sqrt{2}} ln\left(\frac{x_1}{x_2}\right)$
3	0	0	+1	-1	1	1	$b_3 = \sqrt{\frac{1\times 1}{1+1}} \times ln\left[\frac{{x_3}^{1/1}}{{x_4}^{1/1}}\right] = \frac{1}{\sqrt{2}} ln\left(\frac{x_3}{x_4}\right)$

Tabela 3.3. Valores de equilíbrios entre grupos formados em cada etapa de PBS de composições registadas pelo cientista A (*Exemplo 2.3*), em que b_1 corresponde o equilíbrio entre seres vivos e não vivos, b_2 corresponde ao equilíbrio entre animais e vegetais e b_3 corresponde ao equilíbrio entre mineral e água.

	Valores de equilíbrios entre grupos		
Composições	b_1	b_2	b_3
1	-0.5493061	-0.4901291	-1.266965
2	-0.5493061	0.4901291	-1.266965
3	0.4054651	0.000000	0.000000
Média	-0.2310490	0.000000	-0.844643

A média dos valores de equilíbrios entre os grupos apresentados na última linha da Tabela 3.3 indicam-nos que os seres não vivos têm menor peso relativo na composição, os animais e vegetais têm o mesmo peso relativo na composição e, por último, que a água tem maior peso na composição do que os minerais.

CAPÍTULO 4

ANÁLISE EXPLORATÓRIA DE DADOS

4.1. Introdução

Diversas medidas estatísticas permitem sintetizar informações de um conjunto de dados multivariados. As mais comuns são a média e matriz de variâncias-covariâncias. Representações gráficas dos dados também podem ser usados para a visualização de tendências no conjunto de dados. Na análise de dados multivariados de natureza composicional, devemos também ter medidas estatísticas e representações gráficas que permitam descrever numérica e graficamente os dados composicionais. Nesse caso, devemos ter em conta a geometria de seu espaço amostral S^D e, em particular, a distância de Aitchison.

Neste capítulo abordaremos duas estatísticas utilizadas para análise descritiva de dados composicionais e, em seguida, apresentaremos dois tipos de representações gráficas utilizadas na análise desse tipo de dados.

4.2. Descrição numérica

Devido às caraterísticas particulares de dados composicionais, medidas estatísticas usuais da análise multivariada não são muito informativas para esse tipo de dados. Por exemplo, o vetor das médias aritmética e a matriz de variância-covariâncias das partes individuais de uma composição, enquanto medida de tendência central e de dispersão, respetivamente, não são coerentes com a Geometria de Aitchison porque as estatísticas referidas foram definidas de acordo com a geometria euclidiana no espaço real, que não é uma geometria sensível às particularidades de dados composicionais. Duas medidas estatísticas mais usadas para descrição numérica de dados composicionais são o centro e a matriz de variação, que serão definidos a seguir.

Definição 4.1 (Centro)

Seja $\mathbf{X} = [x_{nd}], n = 1, 2, \dots, N, d = 1, 2, \dots, D$, uma amostra aleatória de N composições de D partes. O centro dessa amostra é o vetor de médias geométricas das partes, definido por

$$\boldsymbol{cen}(\mathbf{X}) = C\left[\left(\prod_{n=1}^{N} x_{n1}\right)^{1/N}, \left(\prod_{n=1}^{N} x_{n2}\right)^{1/N}, \dots, \left(\prod_{n=1}^{N} x_{nD}\right)^{1/N}\right], \tag{4.1}$$

em que $C(\cdot)$ é a operação de fecho.

■

Exemplo 4.1 Centro da amostra registada pelo cientista A do Exemplo 2.3

Consideremos novamente a amostra da Tabela 2.1 referente à amostra registada pelo cientista A do *Exemplo 2.3*. O centro dessa amostra é dado por

$$\boldsymbol{cen}(\boldsymbol{A}) = C\left[\left(\prod_{n=1}^{3} x_{n1}\right)^{\frac{1}{3}}, \left(\prod_{n=1}^{3} x_{n2}\right)^{\frac{1}{3}}, \left(\prod_{n=1}^{3} x_{n2}\right)^{\frac{1}{3}}, \left(\prod_{n=1}^{3} x_{n4}\right)^{\frac{1}{3}}\right]$$

$$= C(0.18, 0.18, 0.13, 0.42)$$

$$= (0.20, 0.20, 0.14, 0.46).$$

■

O centro $\boldsymbol{cen}(\mathbf{X})$ é uma medida de tendência central de dados composicionais e corresponde à média aritmética da análise multivariada quando o espaço de resultado é o simplex (Buccianti *et al.,* 2011).

Definição 4.2 (Variância de log-razão)

Seja $\mathbf{x} \in S^D$ uma composição de D partes. A variância de log-razão *(logratio variance)* entre duas partes x_i e x_j de $\mathbf{x}$ é dada por

$$\tau_{ij} = var\left(ln\frac{x_i}{x_j}\right). \tag{4.2}$$

■

A variância de log-razão fornece-nos uma ideia quanto à variabilidade entre duas partes de uma composição. Para termos uma ideia mais abrangente sobre a variabilidade dos dados composicionais temos que calcular variância de log-razão entre todos os pares de partes das composições da amostra, obtendo assim uma matriz de variação, que corresponde à medida de dispersão relativa na análise de dados composicionais (Aitchison, 1986; Pawlowsky-Glahn *et al,* 2015).

Definição 4.3 (Matriz de variação)

Seja $\mathbf{X} = [x_{nd}], n = 1, 2, \dots, N, d = 1, 2, \dots, D$, uma amostra aleatória de N composições de D partes. A matriz de variação *(variation matrix)* de $\mathbf{X}$ é uma matriz quadrada $D \times D$, denotada por $\boldsymbol{T}$, e definida do seguinte modo:

$$\boldsymbol{T} = [\tau_{ij}] = \left[var\left(ln\frac{x_i}{x_j}\right)\right], \qquad i, j = 1, 2, \dots, D \tag{4.3}$$

■

Baseando-se na Definição 4.3 apresentada por Aitchison (1986), Pawlowsky-Glahn *et al* (2015) define também a matriz de variação normalizada para dados composicionais $\boldsymbol{T} = [\tau_{ij}^*]$, em que $\tau_{ij}^* = \frac{1}{\sqrt{2}}\tau_{ij}$ e, supondo a normalidade das log-razões, deduziu o estimador de máxima verosimilhança para a variância de log-razão (4.2) dado por

$$\hat{\tau}_{ij} = \frac{1}{N}\sum_{n=1}^{N}\left(ln\frac{x_{ni}}{x_{nj}} - ln\frac{\hat{g}(\mathbf{x}_i)}{\hat{g}(\mathbf{x}_j)}\right)^2, \tag{4.4}$$

em que $\hat{g}(\mathbf{x}_i)$ e $\hat{g}(\mathbf{x}_j)$ correspondem às médias geométricas dos vetores de partes $\mathbf{x}_j$ e $\mathbf{x}_j$, respetivamente.

Para medir a dispersão global de uma matriz de amostra de dados composicionais $\boldsymbol{X}_{N\times D}$, Pawlosky-Glahn *et al* (2015) definiu uma medida conhecida como variância total (*Sample total variance*) dada por

$$totvar(\mathbf{X}) = \frac{1}{2D}\sum_{i=1}^{D}\sum_{j=1}^{D} var\left(ln\frac{x_i}{x_j}\right) = \frac{1}{2D}\sum_{i=1}^{D}\sum_{j=1}^{D}\tau_{ij}, \tag{4.5}$$

A variância total é, por vezes, chamada de variância métrica (*metric variance*) (Pawlowsky-Glahn *et al,* 2001; 2015).

Para análise completa da variabilidade composicional de uma matriz de dados, Aitchison (1986) considerou uma tabela de variação onde representamos, simultaneamente, as variâncias e as médias log-razões entre as partes das composições da amostra.

Definição 4.5 (Média de log-razão)

Seja $\mathbf{x} \in S^D$ uma composição de D partes. Denota-se por ξ_{ij} a média de log-razão *(logratio mean)* entre duas partes x_i e x_j dada por

$$\xi_{ij} = E\left[ln\frac{x_i}{x_j}\right], \tag{4.6}$$

cujo estimador de máxima verosimilhança para ξ_{ij}, sob o pressuposto de normalidade dos dados, é dado por

$$\hat{\xi}_{ij} = \frac{1}{N}\sum_{n=1}^{N} ln\frac{x_{ni}}{x_{nj}}.$$

■

Definição 4.6 (Tabela de variação)

Seja $\mathbf{x} \in S^D$ uma composição de D partes. A tabela de variação composicional (*variation array*) de $\mathbf{x}$ é dada por

	1	2	3	$\cdots$	$D-1$	D
1	·	τ_{12}	τ_{13}			τ_{1D}
2	ξ_{12}	·	τ_{23}			τ_{2D}
3	ξ_{13}		·			τ_{3D}
$\vdots$	$\vdots$					$\vdots$
$D-1$	ξ_{1D-1}	ξ_{2D-1}	ξ_{3D-1}		·	$\tau_{D-1,D}$
D	ξ_{1D}	ξ_{2D}	ξ_{3D}	ξ_{D-1D}		·

onde os valores ξ_{ij} no triângulo inferior da tabela são as médias de log-razões, sendo o índice da parte no numerador referente ao número da coluna e o da parte no denominador referente ao número da linha da tabela de variação, enquanto que os valores τ_{ij} no triângulo superior são as variâncias de log-razões, sendo o índice da parte no numerador referente ao número da linha e o da parte no denominador referente ao número da coluna da tabela de variação.

■

Para uma melhor visualização da forma como duas partes x_i e x_j da composição $\mathbf{x}$ variam uma em relação a outra, Aitchison (1986) convenientemente seleciona a média e a variância de log-razões simetricamente localizadas em relação à diagonal que separa os triângulos superior e inferior da tabela, cuja interpretação será ilustrada com base no exemplo que se segue.

Exemplo 4.2 Tabela de Variação

Consideremos novamente os dados da Tabela 2.1 referentes a composição do solo registados pelo cientista A referido no *Exemplo 2.3*, onde as partes x_1, x_2, x_3 e x_4 representam, respetivamente,

animal, vegetal, mineral e água. A tabela de variação (Tabela 4.1) permite-nos observar que a maior variação entre duas componentes do solo ocorre entre animal e água, registando $\tau_{14} = 1.26$. As partes vegetal e água também registam o mesmo valor de variação entre elas. Os valores negativos de $\xi_{14} = \xi_{24} = -0.83$ sugerem que, em média, a proporção de água (x_4) no solo é maior do que as proporções de animal (x_1) e de vegetal (x_2). Uma inspeção aos valores da Tabela 2.1 permite-se observar que os dados registados pelo cientista A apoiam esta conclusão. Os menores valores de τ_{ij} ocorrem para as log-razões envolvendo as partes animal e mineral e vegetal e mineral, ou seja, $\tau_{13} = \tau_{23} = 0{,}12$, o que significa que existe menor variação relativa entre animal e mineral e entre vegetal e mineral. Os valores positivos de $\xi_{13} = \xi_{23} = 0{,}37$ indicam que, em média, as proporções de animal e de vegetal no solo são maiores do que a proporção de mineral.

Tabela 4.1. Tabela de variação entre partes da composição do solo registados pelo cientista A

	x_1	x_2	x_3	x_4
x_1	−	0.48	0.12	1.26
x_2	0.00	−	0.12	1.26
x_3	0.37	0.37	−	1.07
x_4	−0.83	− 0.83	− 1.19	−

4.3. Representações gráficas de dados composicionais

Geralmente, para conjuntos de dados composicionais, são usados os seguintes tipos de gráficos: diagramas ternários (gráficos de dispersão fechados de três componentes), gráficos de dispersão de log-razões entre partes, e biplots (gráfico que permite visualizar simultaneamente os indivíduos e as variáveis no mesmo gráfico).

Nesta seção analisaremos apenas os diagramas ternários e os biplots, visto que os gráficos de dispersão de log-razões não são muito informativos para efeito de análise e interpretação de dados composicionais. As ilustrações sobre a interpretação de cada um dos tipos de gráficos abordados serão apresentadas no Capítulo 5, onde aplicaremos as técnicas de análises abordadas a um conjunto de dados do espaço dos codões.

4.3.1. Diagramas ternários

Em Geoquímica, os diagramas ternários constituem uma das principais ferramentas gráficas usadas para a representação de dados composicionais no simplex S^3, sem qualquer transformação aplicada ao conjunto de dados.

A maioria da literatura de análise de dados composicionais (principalmente em Geologia) restringe os gráficos a (sub)composições de três partes porque a representação gráfica de composições com mais do que três partes é mais difícil de visualizar. No caso de $D = 3$ o simplex pode ser representado em $\mathbb{R}^3$ numa superfície triangular, de vértices $A = [k, 0, 0]$, $B = [0, k, 0]$ e $C = [0, 0, k]$, e, geralmente, é visualizado num diagrama ternário no plano $\mathbb{R}^2$, que é uma representação equivalente, conforme representados na Figura 4.1 (a) e Figura 4.1 (b), respetivamente. O valor de k corresponde à constante da operação de fecho na definição do simplex.

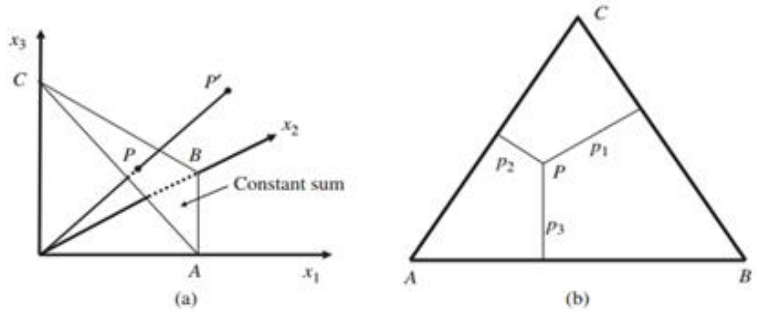


Figura 4.1. (a) Representação do simplex em $\mathbb{R}^3$. (b) Diagrama ternário (Figura extraída de Pawlowsky-Glahn *et al* (2015), pág. 11)

Definição 4.7 (Diagrama ternário)

Um diagrama ternário corresponde a um triângulo equilátero tal que uma amostra genérica $\mathbf{p} = (p_1, p_2, p_3)$ é representada a uma distância p_1 do lado oposto ao vértice A, a uma distância p_2 do lado oposto ao vértice B e a uma distância p_3 do lado oposto ao vértice C.

■

O tripleto (p_1, p_2, p_3) é muitas vezes chamado de coordenadas baricêntricas de p (Boogaart *et al,* 2013).

Para construir um diagrama ternário, começamos por representar os vértices, no sentido contrário ao dos ponteiros do relógio, A, B e C. Assumindo que $A = (u_0, v_0)$ são as coordenadas do vértice A (*Origem*), então, $B = (u_0 + 1, v_0)$ e $C = \left(u_0 + \frac{1}{2}, v_0 + \frac{\sqrt{3}}{2}\right)$, sendo a segunda coordenada do vértice C obtida pelo teorema de Pitágoras. Assim, o diagrama terá a forma apresentada na Figura 4.2. Para representarmos um ponto da amostra com três componentes $\mathbf{x} = (x_1, x_2, x_3)$, fechado para uma constante k, torna-se necessário conhecer as suas coordenadas (u, v), que são obtidas através da combinação linear convexa das coordenadas dos vértices, dada por

$$(u, v) = \frac{1}{k}(x_1 A + x_2 B + x_3 C).$$

Note que as coordenadas da combinação convexa devem ser fechadas para 1, obtidas pela divisão por k.

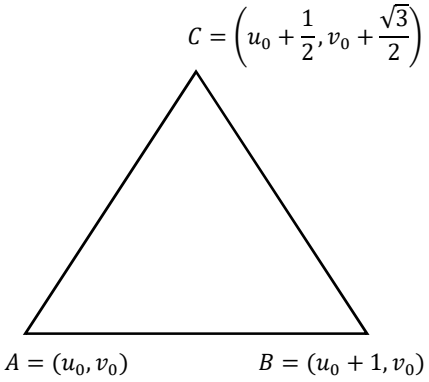


Figura 4.2. Representação de um diagrama ternário, a partir de coordenadas iniciais $(u_0, v_0) = (0.2, 0.2)$.

Boogaart *et al,* (2013) sugere que, para interpretar o diagrama ternário, podemos socorrer da propriedade de que os segmentos ortogonais que ligam um ponto P (ver Figura 4.1 (b)) com os três lados de um triângulo equilátero (as alturas desse ponto) têm soma de seus comprimentos constante: o comprimento de cada segmento é tomada como proporção de uma parte dada. Consequentemente, uma composição representada sobre (ou muito próxima de) uma aresta do triângulo indica a dominância das partes que formam essa aresta, e uma composição representada sobre um vértice indica a dominância da parte associada a esse vértice. Portanto, ao analisar dados composicionais por meio do diagrama ternário, devemos estar atendo aos seguintes padrões:

i. as (sub) composições se concentram num vértice: indica a dominância da parte associada a esse vértice;
ii. as (sub) composições distribuem ao longo de uma aresta: indica a dominância das partes associadas a essa aresta;
iii. as (sub) composições se concentram em torno do baricentro do simplex: indica que as partes representadas têm proporções aproximadamente iguais;
iv. as (sub) composições formam um padrão linear paralelo a um dos lados: indica que as proporções da parte associada ao vértice oposto nas (sub) composições é (aproximadamente) constante;
v. as (sub) composições formam um padrão linear (aproximadamente) perpendicular a um dos lados: indica que as parte associadas a esse lado são (aproximadamente) proporcionais (reduzida variabilidade relativa);
vi. as (sub) composições estiverem dispersas no simplex, indica que as partes apresentam elevada variabilidade relativa entre si.

Os diagramas ternários são especialmente interessantes porque representam os dados tal qual como são: composicional e relativo. As caraterísticas i, ii e iii descritas acima ocorrem, geralmente, quando as três partes das (sub) composições têm valores absolutos muito diferentes entre si. Tal poderá levar os dados a entrar em colapso em um dos vértices (dominância de uma parte), ou ao longo de um dos lados do triângulo (dominância de duas partes), obscurecendo a sua estrutura relativa. Perante essas situações, sugere-se centrar os dados antes de representá-los no diagrama ternário, que, geralmente, exibirá as caraterísticas iv, v ou vi. A centralização dos dados consiste na perturbação de cada linha da matriz de dados, de composições completas, pela inversa do centro, de modo que o conjunto de dados passe a estar distribuído em torno do baricentro do simplex. A realização de uma análise com base nos dados centrados permite uma melhor observação da real tendência no conjunto de dados (Boogaart, K. G. *et al,* 2013; Pawlosky-Glahn, V. *et al,* 2006).

4.3.2. Biplots

O gráfico de dispersão é uma das ferramentas mais utilizadas para a visualização da possível relação entre duas variáveis. O diagrama ternário permite visualizar apenas (sub) composições de três partes. Esses dois tipos de gráfico, no entanto, não permitem a visualização simultânea da possível relação entre mais do que três variáveis (partes), que constitui uma caraterística comum de dados multivariados, incluindo dados composicionais. Uma ferramenta muito popular usada nestes casos é o biplot, introduzido por Gabriel (1971) para dados multivariados e mais tarde, em 2002, adaptados a dados composicionais por Aitchison e Greenacre.

Definição 4.8 (Biplot)

Um biplot é uma representação gráfica, em duas dimensões, da informação contida numa matriz de dados $\boldsymbol{X}_{n\times \boldsymbol{p}}$, em que as n linhas correspondentes às amostras, são representadas como projeção da nuvem dos dados num espaço de duas dimensões e, simultaneamente, sob o mesmo gráfico, são representadas as p colunas da matriz de dados através da projeção dos eixos das variáveis num espaço reduzido.

■

Antes, porém, de construirmos biplots para dados composicionais, apresentamos, de forma sucinta, alguns resultados conhecidos sobre os fundamentos dos biplots.

4.3.2.1. Construção de biplots

Geralmente, a construção de biplots começa com uma transformação da matriz de dados $\boldsymbol{X}_{n\times p}$, de acordo com a natureza dos dados, para que obtenhamos uma matriz transformada, $\boldsymbol{X}_c$, sobre a qual se aplica o biplot. Alguns exemplos dessas transformações são: centralização em relação à média geral, centralização em relação às médias das variáveis, normalização das variáveis, raiz quadrada e transformações log-razões (Aitchison et *al,* 2002). No caso de dados multivariados sem restrições é comum considerarmos a centralização em relação às médias das variáveis, dada por

$$\boldsymbol{X}_c = \boldsymbol{X} - \mathbf{1}\overline{\boldsymbol{X}}^T, \tag{4.7}$$

em que $\mathbf{1}$ é uma matriz $n \times 1$ com todas as entradas iguais a 1 e $\overline{\boldsymbol{X}}$ é um vetor $p \times 1$ que contém as médias de cada uma das colunas de $\boldsymbol{X}_{n\times p}$.

Para contruir o biplot, precisamos de uma factorização da matriz $\boldsymbol{X}_c$ do seguinte modo:

$$\boldsymbol{X}_c = \boldsymbol{G}\boldsymbol{H}^{\boldsymbol{T}}, \tag{4.8}$$

em que $\boldsymbol{G}$ é uma matriz $n \times r$ e $\boldsymbol{H}$ é uma matriz $p \times r$. As linhas de $\boldsymbol{G}$ e as colunas de $\boldsymbol{H}$ fornecem, respetivamente, as coordenadas de n pontos para as linhas e p pontos para as colunas de $\boldsymbol{X}_c$ num espaço euclidiano r-dimensional, chamado espaço completo, cuja dimensão é igual à caraterística de $\boldsymbol{X}_c$. Existem infinitas formas de escolher $\boldsymbol{G}$ e $\boldsymbol{H}$, sendo que certas opções favorecem a representação das linhas e outras a representação das colunas. No entanto, independentemente da escolha de $\boldsymbol{G}$ e $\boldsymbol{H}$, o biplot em r dimensões tem a propriedade de que o produto escalar entre a i-ésima linha de $\boldsymbol{G}$ e a j-ésima coluna de $\boldsymbol{H}$ é igual à entrada (i, j) de $\boldsymbol{X}_c$ (Aitchison *et al*, 2002).

A representação conjunta dos n pontos para as linhas e dos p pontos para as colunas (habitualmente representados através de setas com origem na origem do referencial) corresponde ao biplot exato no espaço completo. No entanto, geralmente, os biplots são representados para dimensões reduzidas da matriz $\boldsymbol{X}_c$, particularmente duas dimensões (i.e., $r = 2$).

A identificação dos fatores $\boldsymbol{G}$ e $\boldsymbol{H}$ em (4.8) pode ser obtida com base na decomposição em valores singulares (*Singular Value Decomposition,* SVD) de $\boldsymbol{X}_c$, dada por

$$\boldsymbol{X}_c = \boldsymbol{U}\mathbf{D}\boldsymbol{V}^T, \tag{4.9}$$

em que $\boldsymbol{U}_{n\times r}$ é a matriz de vetores singulares à esquerda (i.e, vetores próprios de $\boldsymbol{X}_c\boldsymbol{X}_c^T$), $\boldsymbol{V}_{p\times r}$ é a matriz de vetores singulares à direita (i.e, vetores próprios de $\boldsymbol{X}_c^T\boldsymbol{X}_c$) e $\mathbf{D}_{r\times r}$ é a matriz diagonal composta pelos valores singulares positivos (i.e., raízes quadradas dos valores próprios de $\boldsymbol{X}_c^T\boldsymbol{X}_c$ dispostos por ordem decrescente: $d_1 \geq d_2 \geq \cdots \geq d_r > 0$). Pelo teorema de Eckart-Young, podemos

usar os primeiros maiores r^* valores singulares e correspondentes vetores singulares para obter uma matriz $\widehat{\boldsymbol{X}}_c$ de dimensão $n \times p$, que é a melhor aproximação no sentido dos mínimos quadrados de caraterística r^* de $\boldsymbol{X}_c$, ou seja,

$$\left\|\boldsymbol{X}_c - \widehat{\boldsymbol{X}}_c\right\| = \min_{\boldsymbol{Y}} \|\boldsymbol{X}_c - \boldsymbol{Y}\|^2, \tag{4.10}$$

para todas as possíveis matrizes $\boldsymbol{Y}$, de caraterística r^*, em que $\|\cdot\|$ denota a norma matricial de Frobenius[8]. A solução do problema (4.10) é dada por

$$\widehat{\boldsymbol{X}}_c = \boldsymbol{X}_c \boldsymbol{P},$$

em que $\boldsymbol{P}_{p\times p} = \boldsymbol{V}\boldsymbol{V}^T$ e $\boldsymbol{V}_{p\times r^*}$ é uma matriz ortonormal cujas colunas correspondem aos vetores próprios associados aos primeiros e maiores r^* valores próprios da matriz $\boldsymbol{X}_c^T\boldsymbol{X}_c$ (Wedlake, R., 2008).

Para $r^* = 2$, a matriz $\widehat{\boldsymbol{X}}_c$ seria dada por

$$\widehat{\boldsymbol{X}}_c = \begin{pmatrix} u_{11} & u_{21} \\ u_{12} & u_{22} \\ \vdots & \vdots \\ u_{1n} & u_{2n} \end{pmatrix} \begin{pmatrix} d_1 & 0 \\ 0 & d_2 \end{pmatrix} \begin{pmatrix} v_{11} & v_{21} \\ v_{12} & v_{22} \\ \vdots & \vdots \\ v_{1p} & v_{2p} \end{pmatrix}^T. \tag{4.11}$$

O biplot relativo à matriz de dados $\boldsymbol{X}_c$ é construído considerando esta matriz aproximada $\widehat{\boldsymbol{X}}_c$, no espaço reduzido de dimensão $r^* = 2$. A precisão desse biplot corresponde à precisão na aproximação de $\boldsymbol{X}_c$ por $\widehat{\boldsymbol{X}}_c$, e a qualidade da aproximação (4.11) corresponde à proporção da variabilidade explicada (geralmente expressa em percentagem) dada por

$$\pi_r = \frac{d_1^2 + d_2^2}{\sum_{i=1}^{r} d_i^2}. \tag{4.12}$$

Portanto, o SVD fornece-nos uma decomposição adequada para a factorização da matriz $\widehat{\boldsymbol{X}}_c$ conforme apresentada em (4.8) e podemos escolher $\boldsymbol{G} = (d_1^{\alpha}\boldsymbol{u}_1 \quad d_2^{\alpha}\boldsymbol{u}_2)$ e $\boldsymbol{H} = (d_1^{\alpha-1}\boldsymbol{v}_1 \quad d_2^{\alpha-1}\boldsymbol{v}_2)$, resultando

$$\begin{aligned} \widehat{\boldsymbol{X}}_c &= \begin{pmatrix} d_1^{\alpha}u_{11} & d_2^{\alpha}u_{21} \\ d_1^{\alpha}u_{12} & d_2^{\alpha}u_{22} \\ \vdots & \vdots \\ d_1^{\alpha}u_{1n} & d_2^{\alpha}u_{2n} \end{pmatrix} \begin{pmatrix} d_1^{1-\alpha}v_{11} & d_1^{1-\alpha}v_{12} & \cdots & d_1^{1-\alpha}v_{1p} \\ d_2^{1-\alpha}v_{21} & d_2^{1-\alpha}v_{22} & \cdots & d_2^{1-\alpha}v_{2p} \end{pmatrix} \\ &= \begin{pmatrix} \boldsymbol{g_1} \\ \boldsymbol{g_2} \\ \vdots \\ \boldsymbol{g_n} \end{pmatrix} \begin{pmatrix} \boldsymbol{h_1} & \boldsymbol{h_2} & \cdots & \boldsymbol{h_p} \end{pmatrix}, \end{aligned} \tag{4.13}$$

em que $\alpha \in [0, 1]$ é uma constante, chamada parâmetro de forma. Os diferentes valores de α fornecem exatamente a mesma matriz de aproximação e destacará diferentes aspetos da matriz de dados. Existem dois valores particulares de α mais usados na interpretação do biplot, nomeadamente $\alpha = 1$ e $\alpha = 0$, o que significa que os valores singulares são atribuídos completamente para os vetores singulares de $\boldsymbol{U}$ à esquerda ou para os vetores singulares de $\boldsymbol{V}$ à direita, respetivamente. Cada escolha conduz a um biplot com caraterísticas e interpretações diferentes (Greenacre, M., 2010; Wedlake, S., 2008):

[8] $\|A_{m\times n}\| = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n}|a_{ij}|^2}$

1. Se $\alpha = 1$, obtemos linhas nas chamadas coordenadas principais e colunas nas chamadas coordenadas padrão. O biplot resultante é chamado, por alguns autores, de **biplot de forma**, que favorece a representação das observações;
2. Se $\alpha = 0$, obtemos linhas nas chamadas coordenadas padrão e colunas nas chamadas coordenadas principais. O biplot resultante é chamado, por alguns autores, de **biplot de covariância**, que favorece a representação de variáveis.

Soluções alternativas diferem apenas pela alteração de escala ao longo dos eixos horizontal e vertical do biplot. Convencionalmente, as variáveis são representadas através de setas com origem no centro dos dados e as observações são representadas por pontos que correspondem às projeções ortogonais de cada observação sobre o espaço reduzido.

No caso do biplot de covariâncias, que privilegia a representação das variáveis, a factorização (4.13) de $\mathbf{X}_c$ será

$$\mathbf{X}_c = \boldsymbol{GH}^T,$$

em que $\boldsymbol{G} = \boldsymbol{U}$ e $\boldsymbol{H^T} = \boldsymbol{DV}^T$. E, considerando que a matriz de covariâncias de $\boldsymbol{X}_c$ é uma matriz $p \times p$ definida por

$$\boldsymbol{\Sigma} = \frac{1}{n-1}\boldsymbol{X}_c^T\boldsymbol{X}_c, \tag{4.14}$$

as colunas da matriz $\boldsymbol{V}$ em (4.9) também são vetores próprios de $\boldsymbol{\Sigma}$, que pode ser fatorizada como

$$\boldsymbol{\Sigma} = \boldsymbol{V\Lambda V}^T, \tag{4.15}$$

em que $\boldsymbol{\Lambda}$ é uma matriz diagonal $p \times p$ que contém os valores próprios de $\boldsymbol{\Sigma}$ dispostos na ordem decrescente, $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p > 0$, e correspondem aos quadrados dos valores singulares de $\boldsymbol{X}_c$ contidos na matriz $\boldsymbol{D}$, ou seja, $\lambda_i = d_i^2, i = 1, 2, \ldots, p$. Na análise de componentes principais (ACP), cada uma das colunas de $\boldsymbol{V}$ são chamadas de componentes principais (i.e., correspondem aos *loadings* da ACP). Se multiplicarmos ambos os membros da equação (4.9) por $\boldsymbol{V}$, à direita, obtemos

$$\boldsymbol{X}_c\boldsymbol{V} = \boldsymbol{UD} = \boldsymbol{X}_c^*, \tag{4.16}$$

em que $\boldsymbol{X}_c^*$ contém todos os *scores* das componentes principais, o que significa que a matriz $\boldsymbol{V}$ contém também as coordenadas que definem os *scores* de ACP que são representados num biplot de ACP. Do ponto de vista geométrico, o biplot é obtido pela minimização dos quadrados das distâncias entre as observações no espaço p – dimensional e o espaço reduzido r – dimensional (Wdlake, 2008).

De (4.8) podemos reescrever (4.14) do seguinte modo:

$$\begin{aligned}\boldsymbol{\Sigma} &= \frac{1}{n-1}\boldsymbol{X}_c^T\boldsymbol{X}_c \\ &= \left(\frac{1}{\sqrt{n-1}}\boldsymbol{HG}^T\right)\left(\frac{1}{\sqrt{n-1}}\boldsymbol{GH}^T\right).\end{aligned}$$

Considerando que $\boldsymbol{G} = \boldsymbol{U}$, temos que $\boldsymbol{G}^T\boldsymbol{G} = \boldsymbol{U}^T\boldsymbol{U} = \boldsymbol{I_p}$, resulta que

$$\boldsymbol{\Sigma} = \frac{\boldsymbol{HH}^T}{n-1}, \tag{4.17}$$

que corresponde à aproximação de mínimos quadrados da matriz de covariância $\boldsymbol{\Sigma}$ (Aitchison *et al,* 2002). Assim, para que os comprimentos das setas associadas às colunas da matriz correspondam aos valores dos desvios padrão descritos na diagonal de $\boldsymbol{\Sigma}$, devemos tomar em (4.8) a seguinte factorização:

$$\begin{aligned} \boldsymbol{X}_c &= \boldsymbol{G}\boldsymbol{H}^T \\ &= \left(\sqrt{n-1}\boldsymbol{G}\right)\left(\frac{1}{\sqrt{n-1}}\boldsymbol{H}^T\right). \end{aligned}$$

Fazendo $\boldsymbol{G}^* = \sqrt{n-1}\boldsymbol{G}$ e $\boldsymbol{H}^{*^T} = \frac{1}{\sqrt{n-1}}\boldsymbol{H}^T$, resulta que

$$\boldsymbol{X}_c = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^T = \boldsymbol{G}^*\boldsymbol{H}^{*^T}, \tag{4.18}$$

em que $\boldsymbol{G} = \boldsymbol{U}$ e $\boldsymbol{H}^{\boldsymbol{T}} = \boldsymbol{D}\boldsymbol{V}^T$.

Interpretação de biplot de covariância

Dependendo da qualidade da aproximação no biplot da matriz de dados original, podemos interpretar o biplot de covariâncias tendo em conta os seguintes aspetos (Greenacre, M., 2010; Kohler *et al*, 2005):

- Os comprimentos das setas (raios) são estimativas do desvio padrão das respetivas variáveis. Assim, uma seta muito longa indica grande variabilidade da respetiva variável na matriz de dados e vice-versa;
- O cosseno do ângulo formado entre duas setas é uma estimativa da correlação entre as respetivas variáveis. Logo, se o ângulo formado por duas setas for aproximadamente de $90°$, então a correlação entre as variáveies é aproximadamente nula. Por outro lado, se o ângulo formado por duas setas for aproximadamente de $0°$ ou $180°$, então a correlação entre as variáveis em causa é, aproximadamente, 1 ou -1, respetivamente.
- Dado um ponto específico de uma observação (uma linha da matriz de dados), o ponto de interseção da reta que passa por este ponto e perpendicular a uma seta representa o valor desta observação na variável representada pela seta. Assim, pontos de interseção afastados da origem e na direção de uma seta indica valores elevados (superiores à média, que corresponde à origem do referencial) enquanto pontos de interseção na direção oposta ao da seta representa valores abaixos da média da respetiva variável. Se o ponto de interseção estiver na origem, então o valor da observação está próximo da média da respetiva variável.

Por outro lado, no caso do biplot de forma, que privilegia a representação das observações, a factorização (4.13) de $\mathbf{X}_c$ será

$$\mathbf{X}_c = \boldsymbol{G}\boldsymbol{H}^T,$$

em que $\boldsymbol{G} = \boldsymbol{U}\boldsymbol{D}$ **e** $\boldsymbol{H}^{\boldsymbol{T}} = \boldsymbol{V}^T$, com $\boldsymbol{H}^{\boldsymbol{T}}\boldsymbol{H} = \boldsymbol{V}^T\boldsymbol{H} = \boldsymbol{I}_p$**.** E, a matriz a matriz de produtos escalares entre as linhas de $\mathbf{X}_c$ é dada por

$$\begin{aligned} \mathbf{X}_c\mathbf{X}_c^T &= (\boldsymbol{G}\boldsymbol{H}^T)\cdot(\boldsymbol{H}\boldsymbol{G}^T) \\ &= \boldsymbol{G}\cdot(\boldsymbol{V}^T\boldsymbol{V})\cdot\boldsymbol{G}^T = \boldsymbol{G}\boldsymbol{G}^T. \end{aligned}$$

Assim, os produtos escalares e comprimentos dos vetores linha no espaço completo são aproximados otimamente pelo biplot no espaço reduzido. Neste caso, os raios correspondentes às variáveis são ajustados de modo a terem a mesma variância em todas as direções (Aitchison *et al,* 2002).

Interpretação de biplot de forma

Num biplot de forma ($\alpha = 0$), o comprimento de cada seta corresponde à percentagem de variabilidade da respetiva variável. Assim sendo, para um conjunto de dados normalizados, o comprimento da seta de uma variável perfeitamente representada é igual a uma unidade, enquanto que uma variável mal representada tem um raio muito curto. Esta percentagem de variabilidade explicada é chamada de comunalidade (Boogaart, K. G. *et al,* 2013).

Para atingir os objetivos propostos no primeiro capítulo deste trabalho, as propriedades do biplot de forma não são relevantes. Por isso, não debruçaremos muito sobre este tipo de biplot, pelo que, para efeito de análise, recorreremos apenas às propriedades exploratórias do biplot de covariâncias, conforme geralmente ocorre na literatura de Análise Multivariada.

4.3.2.2. Biplot de dados composicionais. Interpretação

Seja $\boldsymbol{X}_{n\times D}$ uma matriz de dados composicionais. Para se representar esse conjunto de dados por meio de biplot, aplicamos inicialmente uma transformação log-razão aos dados antes de centrá-los, de modo que os vetores singulares a esquerda e à direita reproduzam a escala relativa de dados composicionais (Aitchison & Greenacre, 2002). A transformação log-razão usada para construir biplots de dados composicionais é a transformação clr, ou seja, o biplot de dados composicionais é construído sobre uma matriz transformada $\boldsymbol{Z}$ cujas entradas correspondem às coordenadas clr −transformadas calculadas sobre a matriz de dados, que foi previamente centrada em relação às médias das colunas. Assim, na factorização dada na equação (4.13), os vetores $\boldsymbol{g_1}, \boldsymbol{g_2}, \dots, \boldsymbol{g_n}$ são chamados de marcadores de linha de $\widehat{\boldsymbol{Z}}$ e correspondem às projeções das n amostras no plano e, os vetores $\boldsymbol{h_1}, \boldsymbol{h_2}, \dots, \boldsymbol{h_D}$ são chamados de marcadores de colunas de $\widehat{\boldsymbol{Z}}$ e correspondem às projeções das D coordenadas clr no plano (Pawlowsky-Glahn *et al,* 2015).

Na Figura 4.3 está representado um biplot de uma matriz de dados composicionais $\boldsymbol{X}_{n\times D}$, com $D = 4$, onde podemos observar os seguintes elementos:

- Uma origem que representa o centro do conjunto de dados,
- Um vértice para cada uma das D partes (variáveis) em coordenadas clr −transformadas,
- Um ponto como marcador de observações para cada uma das n amostras,
- Um vetor para cada uma das partes, designados por raios.

O segmento de reta que liga dois vértices, por exemplo A e B, $[AB]$, é designado por ligação (*link*).

Interpretação de biplot composicional

As ligações constituem as características básicas de um biplot de covariâncias para dados composicionais, fornecendo as diretrizes para exploração da variabilidade de dados composicionais de acordo com as seguinte propriedades (Boogaart *et* al, 2013; Pawlowsky-Glahn *et al*, 2015):

(a) A ligação entre dois vértices $\boldsymbol{h}_j$ e $\boldsymbol{h}_k$, $[\boldsymbol{h}_j\boldsymbol{h}_k]$, fornece-nos informações sobre a variabilidade da log-razão entre as partes envolvidas, ou seja,

$$\left\|\boldsymbol{h}_j - \boldsymbol{h}_k\right\|^2 \approx \mathrm{var}\left(\ln\frac{\boldsymbol{x}_j}{\boldsymbol{x}_k}\right).$$

Assim, se a qualidade de representação dos dados no biplot for suficientemente elevada,

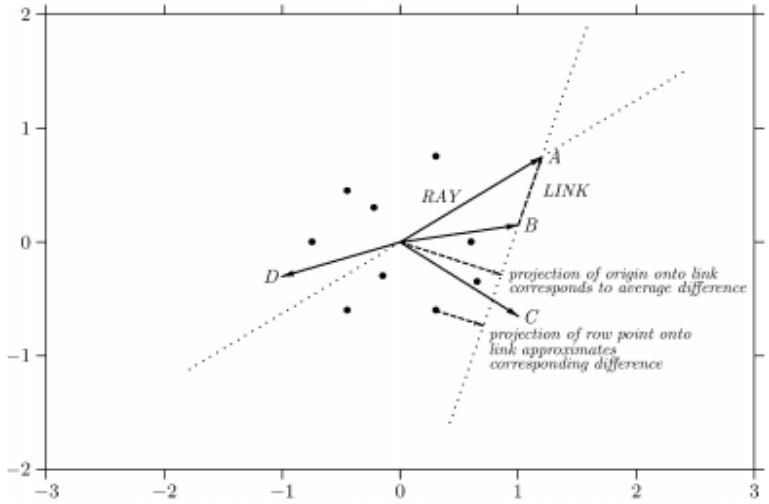


Figura 4.3. Ilustração de um biplot de uma matriz de dados $\boldsymbol{X}_{n\times D}$, sendo: • linhas(amostras); → colunas (partes da composição) (Fonte: Aitchison *et al,* 2002).

- duas variáveis clr −transformadas com ligação muito curta entre si são proporcionais e têm log-razão quase constante (o que corresponde a valores baixos na matriz de variação);
- Inversamente, se a ligação entre duas variáveis clr −transformadas é muito longa, então as partes envolvidas têm uma variabilidade muito grande entre si (entradas elevadas na matriz de variação). Se visualizarmos três setas muito longas a indicarem diferentes direções (formando ângulos de aproximadamente $120°$ entre si), então um diagrama ternário dessas três partes terá dispersão elevada, visto que suas ligações são também muito longas.

(b) O ângulo formado por duas ligações $[\boldsymbol{b}_j\boldsymbol{b}_k]$ e $[\boldsymbol{b}_l\boldsymbol{b}_m]$ fornece-nos informações sobre o valor do coeficiente de correlação entre as duas log-razões,

$$cos\big(\big[\boldsymbol{b}_j\boldsymbol{b}_k\big],[\boldsymbol{b}_l\boldsymbol{b}_m]\big) \approx corr\left(ln\frac{\boldsymbol{x}_j}{\boldsymbol{x}_k}, ln\frac{\boldsymbol{x}_l}{\boldsymbol{x}_m}\right).$$

Assim,

- Se duas ligações formam um ângulo reto entre si significa que as log-razões das partes envolvidas estarão, provavelmente, não correlacionadas;
- Se três ou mais partes colineares têm ligações que formam $0°$ ou $180°$, as log-razões das partes envolvidas estarão perfeitamente correlacionadas (direta ou indiretamente). Neste caso, a subcomposição formada por essas partes deve mostrar um padrão unidimensional de variação, ou seja, essa subcomposição é, aproximadamente, colinear;
- Dois conjuntos de subcomposições colineares, cujas ligações formam ângulos de $90°$ estarão (possivelmente) não correlacionadas.

Para efeito de interpretação, a qualidade do biplot depende da proporção de variância total retida pelo biplot. Qualquer conclusão resultante da análise de biplots pode ser contrastada com outras ferramentas exploratórias dos dados composicionais, como por exemplo a matriz de variação e o diagrama ternário (Boogaart, K. G. *et al*, 2013).

4.3.2.3. Construção de biplot de dados composicionais usando o R

Dado um conjunto de dados composicionais $\mathbf{X}$, em coordenadas originais, é possível construir um biplot de covariâncias de dados composicionais com as funções disponíveis na Biblioteca

`Compositions do` **R**, de duas formas, que são: diretamente, através da função ***princomp()***, ou a partir de SVD da matriz de dados em coordenadas clr-tranformadas.

Construção do biplot através da função ***princomp()***

A construção do biplot composicional, através da função *princomp,* é obtido nos seguintes passos (Boogaart, K. G. *et al* (2013):

1. Fazer **X=acomp(X)** para indicar ao **R** que os dados contidos em **X** são composicionais, devendo aplicar sobre os mesmos a geometria de Aitchison no cálculo de estatísticas;
2. Calcular **princomp(X)**, que retorna um objeto contendo o resultado completo de uma ACP sobre matriz de covariância do conjunto de dados transformados, em coordenadas clr-transformadas, permitindo obter a proporção de variabilidade explicada pelas duas primeiras componentes principais e que corresponde à variabilidade retida pelo biplot dos dados.
3. Aplicar a função **biplot()** sobre o objeto `princomp(`**X**`)` para obter o biplot através da representação das componentes principais armazenadas.

Construção do biplot a partir de SVD da matriz de dados em coordenadas clr-tranformadas

Neste caso, devemos seguir os seguintes passos:

1. Determinar a matriz Z correspondente às coordenadas clr-transformadas de $\boldsymbol{X}$, usando a função **Z=clr(X);**
2. Calcular **svd(Z)** e determinar as matrizes $\boldsymbol{G}^*$ e $\boldsymbol{H}^{*T}$ de acordo com (4.18);
3. Fazer a representação gráfica de duas primeiras colunas da matriz $\boldsymbol{G}^*$ e de duas primeiras colunas da matriz $\boldsymbol{H}^{*T}$, no mesmo referencial, para obter o biplot.

4.3.2.4. Biplot robusto

Muitas vezes, a interpretação de resultados de técnicas estatísticas pode ser prejudicada devido à existência de observações atípicas (*outliers*) no conjunto de dados. Os *outliers* correspondem às observações que apresentam um grande afastamento das restantes ou que são inconsistentes com as demais, e podem ser resultados de erros de medição ou variabilidade inerente dos elementos da população. Assim, a utilização de técnicas estatísticas robustas torna-se particularmente importante porque tais técnicas permitem um bom ajuste aos dados mesmo na presença de *outliers*. E isto é particularmente importante quando lidamos com dados multivariados, como é o caso de dados composicionais (Maronna *et al*, 2006).

A identificação de *outliers* num conjunto de dados exige a consideração de um modelo subjacente ao conjunto de dados. Assim, os *outliers* serão as observações que não são consistentes com o modelo considerado. No caso de dados multivariados, é comum considerar que os dados seguem a distribuição normal multivariada, pelo que assumimos que os *outliers* são os dados oriundos de uma distribuição diferente (Filzmoser *et al*, 2009).

Dada uma matriz de amostra $\boldsymbol{X}_{n\times p}$, a detenção de *outliers* multivariados baseia-se na estimação da estrutura de covariância da matriz dos dados, com o objetivo de medir a distância de cada observação $\boldsymbol{x}_i$ ao centro da nuvem dos dados. Essa distância é calculada com base na métrica de Mahalanobis, definida do seguinte modo:

$$MD(\mathbf{x}_i) = \sqrt{(\mathbf{x}_i - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x}_i - \boldsymbol{\mu})}, \qquad i = 1, 2, \dots, n,$$

em que $\boldsymbol{\mu}$ e $\boldsymbol{\Sigma}$ são, respetivamente, os estimadores robustos da média e da matriz de covariâncias, e $MD(\mathbf{x}_i)$ segue, aproximadamente, a distribuição χ^2_p com p graus de liberdade (Maronna *et al*, 2006). Deste modo, podemos, por exemplo, considerar o quantil de ordem 0.975 de χ^2_p como o valor de corte: observações com valor de MD superior ao valor de corte são consideradas potenciais *outliers* (Filzmoser *et al*, 2009).

As estimativas robustas de $\boldsymbol{\mu}$ e $\boldsymbol{\Sigma}$ podem ser obtidos pelo estimador MCD (*minimum covariance determinant*), que apresenta a vantagem de ser um estimador eficiente e assintoticamente normal (Rousseeuw *et al,* 1999). O estimador MCD carateriza-se pela determinação de um subconjunto de pelo menos h observações cuja matriz de covariância amostral, $\boldsymbol{S}$, tenha o menor determinante. Assim, os estimadores robustos $\widehat{\boldsymbol{\mu}}$ e $\widehat{\boldsymbol{\Sigma}}$ são escolhidos, respetivamente, como a média aritmética e matriz de covariância amostral deste subconjunto, multiplicados por um fator para garantir a consistência dos estimadores sob o pressuposto da normalidade dos dados. A escolha de h determina tanto a robustez como a eficiência dos estimadores, e deve ser, aproximadamente, $h = \frac{3}{4}n$ (Filzmoser *et al,* 2009).

Outliers em dados composicionais

Já vimos que os dados composicionais contêm apenas informação relativa, pelo que somente as razões entre as partes (componentes) são relevantes para a análise. Visto que esses tipos de dados são representados no simplex S^D, eles são transformados (usando transformação alr, clr ou ilr) para o espaço Euclidiano de modo que seja possível a aplicação das usuais técnicas estatísticas desenvolvidos para dados multivariados reais. A transformação clr tem sido uma das mais aplicadas, pelo fato de ela ser coerente com a geometria de Aitchison e permitir uma interpretação dos resultados em termos das variáveis originais. No entanto, a transformação clr resulta em dados colineares, tornando-a inapropriada em técnicas estatísticas robustas baseadas na matriz de covariância, como é o caso dos estimadores MCD para $\boldsymbol{\mu}$ e $\boldsymbol{\Sigma}$, que só podem ser determinados para conjunto de dados não singulares, cuja caraterística da matriz seja igual ao número de variáveis (Filzmoser *et al*, 2009; Maronna *et al*, 2006). Por outro lado, a transformação ilr não apresenta o problema de colinearidade e goza de propriedades compatíveis com qualquer tipo de análise estatística no espaço Euclidiano.

Definição 4.9 (Distribuição Normal de dados composicionais)

Seja $\boldsymbol{z} = (z_1, z_2, \dots, z_{D-1})$ as coordenadas ilr-transformadas de uma composição $\mathbf{x} \in S^D$. Dizemos que um conjunto de dados composicionais $\boldsymbol{X}_{n\times D}$ tem distribuição normal no simplex se $\boldsymbol{Z}_{n\times(D-1)} = ilr(\boldsymbol{X})$ tem distribuição normal multivariada em $\mathbb{R}^{D-1}$. Neste caso, escrevemos $\boldsymbol{Z} \sim N(\boldsymbol{\mu_z}, \boldsymbol{\Sigma_z})$ e $\boldsymbol{X} \sim N_{S^D}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. ■

A matriz de covariâncias $\boldsymbol{\Sigma_z}$ é não singular, ou seja, é positiva definida, com $|\boldsymbol{\Sigma_z}| \neq 0$ e existe $\boldsymbol{\Sigma_z}^{-1}$ (Pawlowsky-Glahn *et al*, 2015). Assim, sob o pressuposto de normalidade multivariada dos dados no simplex, a distância de Mahalanobis $MD(\boldsymbol{z_i}), i = 1, 2, \dots, n$, segue a distribuição χ^2_{D-1}, e o quantil de ordem $0{,}975$ pode ser usado como valor de corte para separar as observações regulares daquelas que constituem potenciais *outliers* (Filzmoser *et al*, 2012).

Pretende-se que potenciais *outliers* possam ser identificados por meio de representações gráficas dos dados. No caso de dados multivariados como é o caso de dados composicionais, o biplot poderia

constituir uma ferramenta adequada para tal, pois permite a visualização de padrões na estrutura de dados multivariados, no espaço reduzido 2-dimensional.

Filzmoser *et al* (2009) propôs o uso de biplots robustos para lidar e identificar *outliers* em dados composicionais. Assim, para a realização de APC e construção de biplots robustos aqueles autores sugeriram considerar dados em coordenadas ilr −transformadas, relativamente a uma dada base, com vista à obtenção dos *loadings* e dos *scores* robustos. Mas, para a interpretação da ACP realizada sobre dados em coordenadas ilr-transformadas, sugerem que voltemos a transformar os dados para o espaço de dados em coordenadas clr-transformadas, onde a interpretação do biplot é conhecida.

Seguindo abordagem proposta por Filzmoser *et al* (2009), consideremos uma amostra de dados composicionais $\boldsymbol{X}_{n\times D}$ e a correspondente matriz em coordenadas ilr-transformadas $\boldsymbol{Z}_{n\times(D-1)}$, de valor médio $\boldsymbol{\mu}_Z$ e matriz de covariância $\boldsymbol{\Sigma_Z}$ e com estimativas robustas obtidas pelos estimadores MCD $\widehat{\boldsymbol{\mu}}_{\boldsymbol{Z}}$ e $\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{Z}}$, respetivamente. Tomando a SVD de $\widehat{\boldsymbol{\Sigma}}_{\mathbf{z}}$, isto é, $\widehat{\boldsymbol{\Sigma}}_{\mathbf{z}} = \boldsymbol{V_Z}\boldsymbol{\Lambda_Z}\boldsymbol{V_Z^T}$**,** então a matriz dos *scores* será a matriz $\boldsymbol{Z}^*_{n\times(D-1)}$, a qual descreve os dados $\boldsymbol{Z}$ centrados no espaço das componentes principais robustas, ou seja,

$$\boldsymbol{Z}^* = \left[\boldsymbol{Z} - \mathbf{1}\widehat{\boldsymbol{\mu}}_{\boldsymbol{Z}}^T\right]\boldsymbol{V_Z^T}, \tag{4.21}$$

em que $\mathbf{1}$ é um vetor n-dimensional de entradas iguais à unidade e $\boldsymbol{V_Z}$ é a matriz dos *loadings,* cujas colunas contêm os vetores próprios de $\widehat{\boldsymbol{\Sigma}}_{\mathbf{z}}$**.**

Se a matriz de dados original, $\boldsymbol{X}_{n\times D}$ tiver caraterística D, a matriz $\boldsymbol{Z}$ terá caraterística completa $D-1$, e o estimador MCD poderá ser usado para obter as estimativas robustas de $\widehat{\boldsymbol{\mu}}$ e $\widehat{\boldsymbol{\Sigma}}$, resultando em componentes principais robustas contidas na matriz $\boldsymbol{V_Z}$ e matriz de *scores* $\boldsymbol{Z}^*$**.**

Para a interpretação dos *loadings* e *scores* robustos (4.21) o biplot robusto devemos representar aqueles *loadings* e *scores* em suas respetivas coordenadas clr-transformadas. Consequentemente, usando (2.29) a matriz de *scores* robustos em coordenadas clr-transformadas será dada por

$$\boldsymbol{Y}^* = \boldsymbol{Z}^*\boldsymbol{\Psi}^T, \tag{4.22}$$

Analogamente, temos que

$$\begin{aligned}\widehat{\boldsymbol{\Sigma}}_{\mathbf{Y}} = \widehat{\boldsymbol{\Sigma}}_{\boldsymbol{Z\Psi'}} &= \boldsymbol{\Psi}\cdot\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{Z}}\cdot\boldsymbol{\Psi'}\\ &= \boldsymbol{\Psi V_Z}\boldsymbol{\Lambda_Z}\boldsymbol{V_Z^T}\boldsymbol{\Psi^T}\\ &= \boldsymbol{V_Y}\boldsymbol{\Lambda_Z}\boldsymbol{V_Y^T},\end{aligned} \tag{4.23}$$

pelo que $\boldsymbol{V_Y}$ corresponde à matriz dos *loadings* robustos em coordenadas clr-transformadas. Observemos também que, devido à relação de linearidade entre as transformações ilr e clr-transformadas, os valores próprios não nulos de $\widehat{\boldsymbol{\Sigma}}_{\mathbf{z}}$ são iguais aos de $\widehat{\boldsymbol{\Sigma}}_{\mathbf{Y}}$, pelo que a percentagem de variabilidade explicada contida na diagonal da matriz $\boldsymbol{\Lambda_Z}$ é a mesma para a correspondente matriz $\boldsymbol{\Lambda_Y}$ em coordenadas clr-transformadas (Filzmoser *et al,* 2009).

Agora, para construir o biplot composicional robusto para uma matriz de dados coordenadas clr-tranformadas $\boldsymbol{Y}$, precisamos fatorizar $\boldsymbol{Y}$ na forma $\boldsymbol{Y} = \boldsymbol{G_Y}\boldsymbol{H_Y^T}$, de acordo com (4.8), com base nas matrizes de *loadings* e de *scores* robustos $\boldsymbol{Y}^*$ e $\boldsymbol{V_Y}$. Para tal, consideremos a SVD de $\boldsymbol{Y}$, isto é, $\boldsymbol{Y} = \boldsymbol{U_Y}\boldsymbol{D_Y}\boldsymbol{V_Y^T}$, em que $\boldsymbol{D_Y} = \boldsymbol{\Lambda_Y^{1/2}}$, isto é, $\boldsymbol{D_Y}$ é a matriz diagonal cujas entradas correspondem às raízes quadradas dos elemento de $\boldsymbol{\Lambda_Y}$**.** Então, de (4.16) temos que

$$\boldsymbol{Y}^* = \boldsymbol{U_Y}\boldsymbol{D_Y} \tag{2.24}$$

E, multiplicando ambos os membros de (2.24) por $\boldsymbol{D_Y^{-1}}$, à direita, resulta que

$$\boldsymbol{Y^* D_Y^{-1} = U_Y}. \tag{2.25}$$

De (4.25) e usando (4.13), o biplot de covariâncias composicional robusto em coordenadas clr-transformadas é obtido escolhendo-se $\boldsymbol{G_Y = Y^* D_Y^{-1}}$ e $\boldsymbol{H_Y^T = D_Y V_Y^T}$.

CAPÍTULO 5

APLICAÇÃO AO ESPAÇO DOS CODÕES

Neste capítulo consideraremos um conjunto de dados do espaço dos codões, constituído pelas 31 espécies listadas na Tabela 1.1, onde exploraremos a variação relativa das frequências dos nucleótidos, considerando diferentes situações (4 casos de estudos) usando técnicas de análise de dados composicionais abordadas nos capítulos 2 a 4. Em cada caso, complementaremos o estudo com análise na perspetiva absoluta.

Takeuchi *et al* (2003) analisou um conjunto de dados do espaço dos codões, constituído por 27 espécies, através de técnicas estatística multivariada, sem considerar a natureza composicional dos dados. Neste capítulo, procuramos analisar o nosso conjunto de dados do espaço dos codões também através das técnicas de análise de dados composicionais abordadas ao longo deste trabalho. Aplicaremos o biplot de covariância sobre dados em coordenadas originais (dados brutos) e sobre dados em coordenadas log-razões transformadas (clr e ilr). Devido a relação linear existente entre coordenadas clr e ilr-transformadas (Eq. 2.18 e Eq. 2.32), os pontos e as setas dos biplots composicionais nessas duas coordenadas exibem os mesmos padrões, e estão sujeitas às mesmas interpretações (no espaço de coordenadas clr-transformadas). Para os dados em cada uma das coordenadas referidas, compararemos os resultados obtidos por meio do biplot clássico e do biplot robusto, sendo que este último permite contornar eventuais distorções dos resultados causados pela presença de *outliers* no conjunto de dados (Filzmoser *et al*, 2009).

5.1. Métodos de análise dos dados

O nosso conjunto de dados contém a composição de bases de nucleótidos de 31 espécies pertencentes aos cinco reinos de seres vivos, sendo: 16 animais, 4 plantas, 5 bactérias, 3 fungos e 3 protozoários. Utilizaremos os biplots tradicional (para dados em bruto) e composicional para explorar informação absoluta e relativa contida neste conjunto de dados do espaço dos codões. Os biplots serão aplicados ao conjunto de dados de quatro maneiras diferentes, conforme se segue:

Estudo 1: *Frequências relativas das bases em cada posição dos codões, de forma separada.*

Neste caso, trataremos as quatro bases em cada uma das três posições dos codões como um conjunto de dados específico, sendo cada um formado por 4 variáveis e 31 observações. Assim, o primeiro conjunto de dados corresponde apenas às frequências das bases na primeira posição dos codões (i.e., x_1, x_2, x_3, x_4), o segundo conjunto corresponde às frequências das bases na segunda posição dos codões (i.e., x_5, x_6, x_7, x_8) e, por fim, o terceiro conjunto de dados corresponderá às frequências das bases que ocupam a terceira posição dos codões. Notemos que, embora o número total de bases nas três posições seja o mesmo, as bases apresentam diferentes frequências em cada posição, ou seja, algumas bases privilegiam umas posições mais do que as outras.

Estudo 2: *Frequências relativas das bases nas três posições do codão, de forma conjunta*

Neste caso, embora consideremos as frequências relativas das bases em cada posição dos codões, analisaremos as frequências nas três posições como se fosse apenas um conjunto de dados, contendo 12 variáveis e 31 observações, sendo

$$x_1 + x_2 + x_3 + x_4 + x_5 + x_6 + x_7 + x_8 + x_9 + x_{10} + x_{11} + x_{12} = 3. \quad (5.18)$$

Do ponto de vista composicional, o Estudo 1 corresponde ao estudo de subcomposições da composição completa considerada no Estudo 2. Assim, as conclusões obtidas nestas duas análises deverão ser coerentes.

Estudo 3: *Análise de dados fundidos – soma das frequências de cada uma das bases*

Neste caso, analisaremos a fusão $\boldsymbol{a} = (a_1, a_2, a_3, a_4)$, em que cada uma das componentes de $\boldsymbol{a}$ corresponde à soma das frequências de cada uma das quatro bases nas três posições do codão, ou seja,

$$a_1 = A1 + A2 + A3 = x_1 + x_5 + x_9,$$
$$a_2 = C1 + C2 + C3 = x_2 + x_6 + x_{10},$$
$$a_3 = G1 + G2 + G3 = x_3 + x_7 + x_{11},$$
$$a_4 = T1 + T2 + T3 = x_4 + x_8 + x_{12}.$$

A fusão $\boldsymbol{a}$ assim definida visa analisar as proporções de cada uma das bases tendo em conta as suas frequências nas três posições dos codões. Deste modo, a composição original de 12 componentes fica reduzida a uma composição de apenas 4 componentes.

Estudo 4: *Análise de dados fundidos – análise em termos do teor C+G e A+T nas três posições dos codões*

Considerando os pares formados na cadeia de ADN, pretendemos com esta fusão analisar as espécies em termos de frequências das bases de cada um desses pares. Para isso, consideramos uma fusão $\boldsymbol{a} = (a_1, a_2, a_3, a_4, a_5, a_6)$, em que $a_1 = A1 + T1, a_2 = C1 + G1, a_3 = A2 + T2, a_4 = C2 + G2, a_5 = A3 + T3$ e $a_6 = C3 + G3$. Deste modo, a nossa análise será feita sobre uma composição de dimensão reduzida com 6 componentes.

Os biplots de covariâncias para os dados originais e os biplots composicionais serão usados como forma de complementar as conclusões que se pode extrair pela análise de cada um. Tendo em conta o conhecimento do reino a que cada uma das 31 espécies pertencem, iremos analisar a capacidade discriminativa das espécies pelos diferentes biplots considerados. Em particular, destacaremos se os pontos nos biplots estão agrupados por reinos. Para tal, na construção de cada biplot, pontos (espécies) referentes a cada um dos reinos são identificados por diferentes cores: preto para animais, azul para plantas, verde para protozoários, magenta para bactérias, e vermelho para fungos.

5.2. Resultados

Estudo 1: *Frequências relativas das bases em cada posição dos codões, de forma separada.*

Na Figura 5.1. podemos observar biplots clássicos construídos para os conjuntos de dados referentes às frequências de cada base fixando a sua posição (primeira, segunda ou terceira) nos codões. Para cada posição, construímos três biplots, respetivamente, para dados brutos (i.e., dados originais), dados em coordenadas clr-transformadas e dados em coordenadas ilr-transformadas.

Quanto às frequências de bases na primeira posição dos codões, na Figura 5.1 (a) observamos o biplot clássico construído a partir dos dados originais, onde não consideramos a natureza composicional dos dados. Na Figura 5.1. (b) e (c) estão biplots composicionais, em coordenadas clr e ilr-transformadas,

respetivamente. Nos três biplots podemos observar a formação de grupos no espaço reduzido, com uma nítida separação entre animais e bactérias, pela segunda componente principal. No biplot para dados em originais (Figura 5.1(a)), observamos que as espécies pertencentes ao reino animal tendem a mostrar frequência do nucleótido C na primeira posição acima da média aritmética, definindo assim um grupo coeso, com exceção de *Ce* (nº 22) e Am (nº 31), que se encontram mais dispersos, ambos com frequências do nucleótido C abaixo da média. Este mesmo comportamento das espécies pertences ao reino dos animais é também observado nos biplots referentes às bases em cada uma das restantes posições dos codões (Figura 5.1 (d) e (g)), onde apresentam frequências do par (C, G) acima da média, em oposição às frequências do par (A, T) que aparecem abaixo da média (Figura 5.1 (d) e (g)).

Os biplots para dados em variáveis originais realçam um contraste entre animais e bactérias, no que diz respeito às frequências de nucleótidos nas três posições de seus codões. De facto, enquanto os animais favorecem o par (C,G), as bactérias tendem a favorecer o par (A, T) nos seus codões. Por exemplo, ao contrário da classe animal, as bactérias tendem a mostrar frequências do nucleótido C na primeira posição (assim como nas restantes) abaixo da média, com exceção de *Ec* (nº 12). Além disso, verificamos que na primeira posição dos codões as bactérias favorecem o nucleótido G, e na segunda e terceira posições as espécies desta classe tendem a mostrar frequências do par (C, G) abaixo da média.

Em relação às espécies pertencentes ao reino das plantas, verificamos que as plantas observadas tendem a mostrar dominância do nucleótido T nas três posições dos codões, com exceção da espécie *Os* (nº 8) que apresenta a frequência do nucleótido T abaixo da média, favorecendo, por sua vez, as bases do par (C, G). Entretanto, dado o reduzido número de plantas incluídas na análise, os padrões de frequências de nucleótidos das espécies deste reino, observados nos biplots, não podem ser considerados conclusivos. O mesmo se dá em relação aos fungos e protozoários, cujos números de observações incluídas na análise foram apenas de 3 para cada uma dessas classesiºo~+. Além disso, as distâncias entres as observações dessas últimas são muito elevadas entre si, não se verificando a formação de grupos.

As percentagens de variabilidade de dados retidas pelas duas primeiras componentes principais representadas nos biplots, para as variáveis originais, foram muito boas 96,2%, 96,7% e 99,1%, respetivamente, para bases na primeira, segunda e terceira posição). Assim, os comprimentos das setas nos biplots (Figura 5.1 (a), (d) e (g)) nos fornecem uma boa ideia sobre o padrão de variação das frequências de bases em cada uma das posições dos codões. Por exemplo, na primeira posição, as bases A e C são as que apresentam maiores desvios (com valores relativamente iguais), enquanto T é a base que apresenta menor desvio (Figura 5.1 (a)). A base que apresenta maior desvio na segunda posição é a base A. (Figura 5.1 (d)). Mas, em comparação com as frequências de bases nas restantes posições, verifica-se maior regularidade nas frequências de base nesta posição, enquanto as bases da terceira posição apresentam maiores desvios, com destaque para o nucleótido C (Figura 5.1 (g)).

A direção das setas nos biplots para dados originais fornecem-nos informações sobre a correlação entre as bases. Por exemplo, na Figura 5.1. (a), observamos que, na primeira posição dos codões, as bases A e T estão fortemente correlacionadas (i.e, $corr(A,T) \approx 1$), em oposição ao grupo formado pelas bases C e G. Além disso, verificamos que existe uma correlação negativa entre os pares (A, T) e (C, G). Esses dois padrões de correlação entre as bases devem-se, provavelmente, ao facto de que, teoricamente, as frequências de bases pertencentes a cada um dos pares serem as mesmas, e que um incremento nas frequências de bases de um dos pares implica redução das frequências de bases

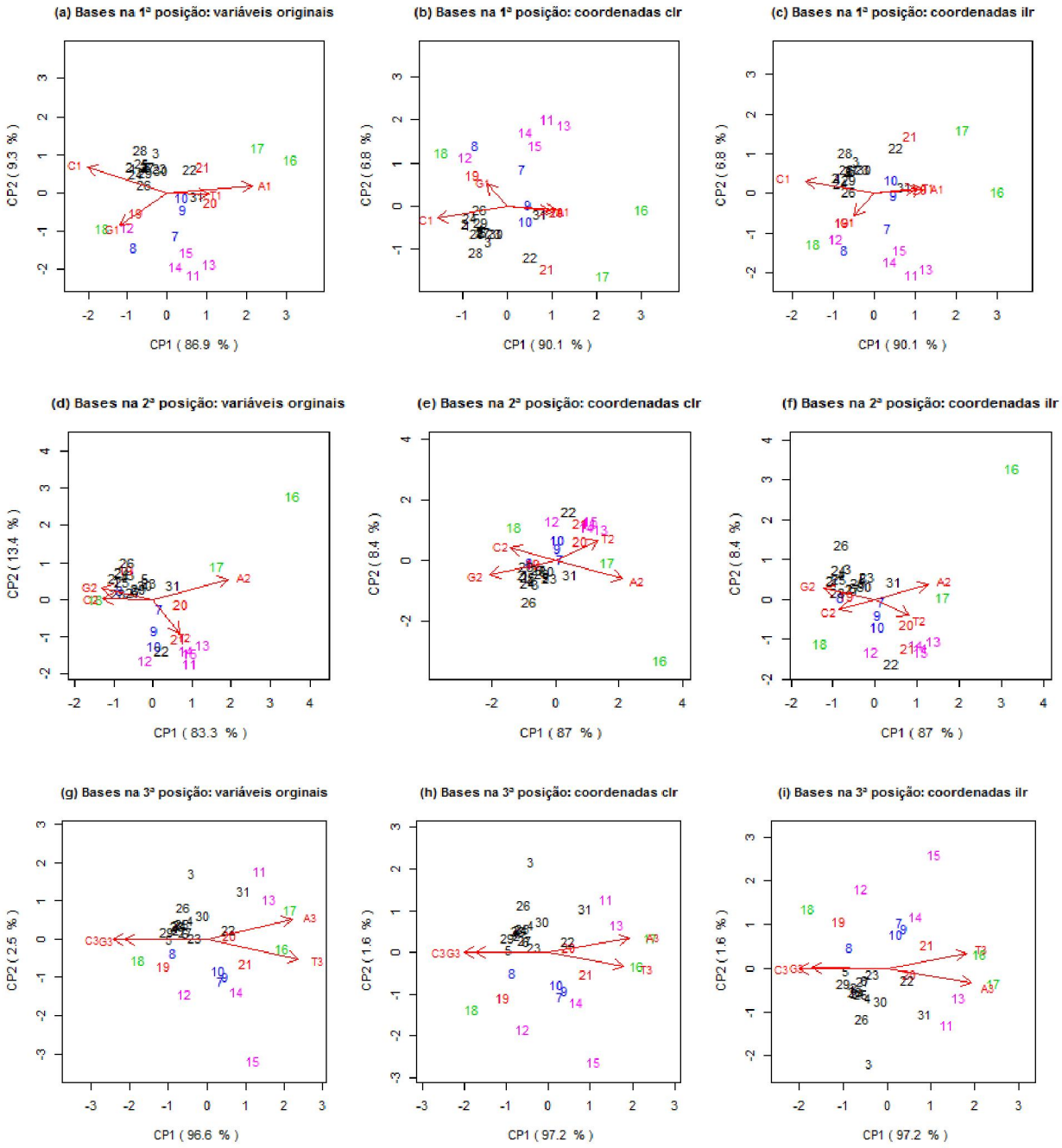


Figura 5.1. Biplots clássicos construídos a partir de dados originais, dados em coordenas clr —transformadas e dados em coordenadas ilr —transformadas (da esquerda para direita), referentes às frequências de bases na primeira, segunda e terceira posições dos codões (de cima para baixo).

do par oposto (Lei de Chargaff). Na segunda posição as bases do par (C, G) estão fortemente correlacionadas, em oposição às bases do par (A, T), com as quais evidenciam correlação negativa (Figura 5.1 (d)). Quanto às bases na terceira posição, podemos observar uma perfeita correlação positiva entre as bases do par (C, G), em oposição às bases do par (A, T), que também evidenciam uma correlação positiva não fraca entre si. Além disso, existe uma forte correlação negativa entre esses dois pares de bases, assim como acontece em relação às bases na primeira e segunda posição dos

codões. Para apoiar estas conclusões, apresentamos nas Tabelas 5.1 e 5.2 valores referentes aos desvios e correlações entres as bases, em cada uma das três posições dos codões.

Tabela 5.1. Valores dos desvios padrão de frequências das bases em cada umas das três posições do codão, que reforçam as conclusões obtidas pela análise do padrão de variabilidade dos dados exibido pelos biplots para dados em coordenadas originais, apresentados nas Figuras 5.1 (a), (d) e (g), onde podemos ver que as bases da terceira posição apresentam valores de desvios mais elevados.

Posições do codão	Primeira posição				Segunda posição				Terceira posição			
Bases	A	C	G	T	A	C	G	T	A	C	G	T
Variáveis	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}	x_{12}
Valores dos desvios	0.046	0.045	0.031	0.025	0.045	0.030	0.031	0.025	0.075	0.081	0.072	0.080

Tabela 5.2. Valores de correlações entre frequências de bases em cada uma das posições do codão, que reforçam as conclusões obtidas pela análise dos padrões de correlação entre frequências de bases em cada posição do codão exibido pelas setas dos biplots para dados em coordenadas originais, apresentados nas Figuras 5.1 (a), (d) e (g), onde podemos observar uma tendência de forte correlação positiva entre as bases dos pares (A, T) e (C, G), e forte correlação negativa entre esses pares de bases.

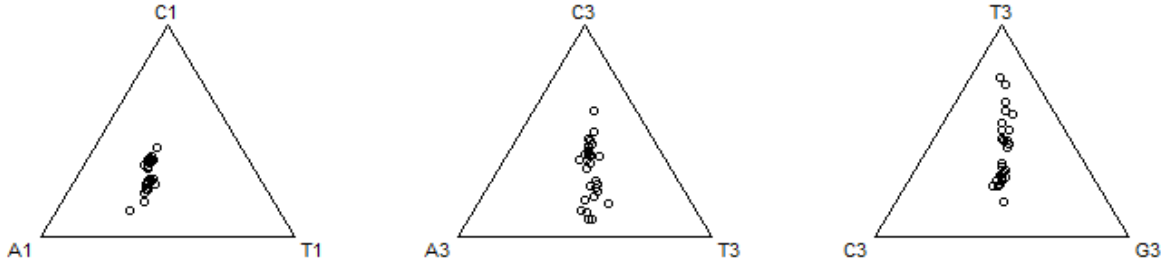
Posições do codão	Primeira posição				Segunda posição				Terceira posição			
Bases	A	C	G	T	A	C	G	T	A	C	G	T
A	1.00	−0.90	−0.83	0.78	1.00	−0.90	−0.85	0.35	1.00	−0.96	−0.96	0.90
C		1.00	0.59	−0.85		1.00	0.83	−0.62		1.00	0.97	−0.97
G			1.00	−0.75			1.00	−0.72			1.00	−0.97
T				1.00				1.00				1.00

Por outro lado, as ligações (*links*) entre as setas que representam as variáveis nos biplots composicionais clássicos fornecem informações sobre a variação relativa entre as bases. Por exemplo, na Figura 5.1. (b) e (c), referente às bases na primeira posição, podemos observar que a ligação entre A e T é muito curta, o que indica que a log-razão $ln(A/T)$ é quase constante e, portanto, que as frequências das bases A e T são proporcionais. Adicionalmente, visto que as bases do par (A, T) preservam ligações muito longas com as bases do par (C, G), significa que log-razões envolvendo bases desses dois pares apresentam elevada variabilidade, principalmente as log-razões $ln(A/C)$ e $ln(T/C)$. Um diagrama ternário de dispersão entre as bases do conjunto {A, C, T} deverá exibir apenas uma variabilidade unidimensional, isto é, as observações exibirão um padrão linear de variação, conforme podemos confirmar na Figura 5.2. (a). Quanto aos biplots composicionais para bases na segunda posição, Figuras 5.1. (e) e (f), verificamos que as ligações [A2, G2] e [T2, C2] são aproximadamente paralelas, pelo que as log-razões $ln\frac{A2}{G2}$ e $ln\frac{T2}{C2}$ estão fortemente correlacionados. Por fim, nas Figuras 5.1. (h) e (i), podemos observar que as bases C e G preservam uma ligação muito curta entre si, o que significa que essas bases têm log-razões constante, indicando que as frequências dessas bases são proporcionais entre si. Essas conclusões sobre a variabilidade relativa entre as bases de cada posição dos codões são reforçadas na Tabela 5.3, onde podemos observar valores dos triângulos superiores da tabela de variação de log-razões (Definição 4.6) entre as bases de cada uma

Tabela 5.3. Triângulos superiores de tabelas de variação de log-razões entre frequências de bases em cada uma das posições dos codões, que reforçam as conclusões obtidas pela análise dos padrões exibidos pelas setas dos biplots composicionais apresentados nas Figuras 5.1 (b), (c), (e), (f), (h) e (i), onde podemos observar uma tendência das log-razões envolvendo bases de cada um dos pares (A, T) e (C, G) apresentar variabilidades reduzidas, enquanto log-razões envolvendo bases de distintos pares apresentam variabilidades mais elevadas, exceto para as bases da segunda posição, que apresentam o mesmo padrão de reduzida variabilidade relativa entre todas as bases.

Posições do codão	Primeira posição				Segunda posição				Terceira posição			
Bases	A	C	G	T	A	C	G	T	A	C	G	T
A	–	0.144	0.058	0.007	–	0.073	0.095	0.014	–	0.518	0.447	0.016
C		–	0.037	0.128		–	0.011	0.046		–	0.014	0.488
G			–	0.049			–	0.069			–	0.419
T				–				–				–

das três posições dos codões. Além disso, o facto de que as setas que representam as bases dos pares (A, T) e (C, G) serem, aproximadamente, colineares, indica que um diagrama ternário envolvendo quaisquer três bases da terceira posição dos codões tenderá a exibir um padrão de variação aproximadamente linear (Figura 5.2.(b) e (c)).



(a) (b) (c)

Figura 5.2. Diagramas ternários que mostram os padrões de variação linear exibidos pelas bases dos grupos {A1, C1, T1}, {A3, C3, T3}, e {C3, G3, T3}, cujas setas das bases em cada grupo eram aproximadamente colineares nos biplots composicionais referentes às frequências de bases na primeira e terceira posição do codão, representados nas Figuras 5.1 (b), (c), (h) e (i).

Estudo 2: *Frequências relativas das bases nas três posições dos codões, de forma conjunta*

Na Figura 5.3 representamos biplots de covariâncias construídos para dados referentes às frequências de bases nas três posições dos codões, onde observamos que os padrões exibidos pelas variáveis (bases) e pelas observações têm muita semelhança com os padrões já observados na Figura 5.1. quando analisámos as bases em cada uma das posições separadamente. No biplot clássico para dados em coordenadas originais (Figura 5.3 (a)) é possível visualizar que a segunda componente principal estabelece uma clara divisão entre espécies pertencentes ao reino Monera (bactérias, representadas pela cor magenta) e Animais (animais, representados pela cor preta). As observações referentes às espécies do reino Protista (os protozoários) estão dispersas no espaço reduzido definido pelas duas primeiras componentes principais. Assim, as observações referentes aos protozoários parecem constituir eventuais *outliers* no conjunto de dados em análise. As espécies do reino Animal (com exceção de duas espécies) apresentam frequências dos nucleótidos C e G acima da média em todas as posições dos codões (com exceção do nucleótido G da primeira posição), enquanto apresentam frequência abaixo da média para as bases A (na terceira posição) e T (nas três posições). Verificamos

que os animais designados por *Ce* (nº 22) e *Am* (nº 31) apresentam um padrão de frequências de bases diferentes das restantes espécies do reino a que pertencem, visto que favorecem os nucleótidos A e T nos seus codões, apresentando, por sua vez, as frequências das bases C e G abaixo da média nas três posições de seus codões. Destacamos ainda as bactérias em oposição aos animais, pois, embora as bactérias favoreçam o nucleótido G na primeira posição de seus codões, as bactérias observadas exibem frequências dos nucleótidos C e G abaixo da média, e frequências dos nucleótidos A (na terceira posição) e T (nas três posições) acima da média. Em relação às plantas, observamos que as espécies observadas tendem a favorecer apenas uma base na primeira posição de seus codões (nucleótido G) e outra base na segunda posição (nucleótido T), com exceção da espécie *Os* (nº 8) que apresenta frequências das bases C e G acima da média em todas as posições dos seus codões.

A percentagem de variabilidade de dados retida pelas duas primeiras componentes principais é consideravelmente boa (94,7%), pelo que o padrão de variação de dados apresentado no biplot para dados originais é fiável. Assim, na Figura 5.3 (a), observamos que as bases na terceira posição são as que apresentam maiores valores de desvios em relação à média. Na Tabela 5.4 podemos ver os valores exatos dos desvios padrão de cada uma das bases das três posições do codão.

Uma notável vantagem da análise conjunta das bases nas três posições dos codões relaciona-se com o facto podermos visualizar possíveis relações entre bases localizados em diferentes posições dos codões. Por exemplo, na Figura 5.3 (a), podemos observar que, em alguns casos, a frequência de uma dada base numa posição está fortemente correlacionada com a sua frequência noutras posições dos codões, nomeadamente as bases pertencentes aos grupos {A1, A2}, {C1, C2, C3}, {G2, G3} e {T1, T3}. Adicionalmente, verificamos que existe também uma correlação positiva entre algumas bases da primeira posição dos codões com o correspondente par de ligação da terceira posição dos codões, nomeadamente, as bases dos pares (T1, A3) e (C1, G3). Por outro lado, assim como observamos nos biplots para dados originais, onde consideramos apenas frequências de bases em cada uma das três posições dos codões, no presente biplot (Figura 5.3 (a)), verificamos também a separação das variáveis em dois grupos, pela primeira componente principal. Um grupo é formado pelas bases do par (A, T), em oposição ao grupo formado pelas bases do par (C, G), onde as bases dentro de cada grupo tendem a estar fortemente correlacionadas entre si, mas negativamente correlacionadas com as bases pertencentes ao grupo oposto, conforme acontece, por exemplo, com as bases pertencentes aos grupos {A3, T1, T3} e {C1, C3, G2, G3}. Procedendo a uma ampliação do biplot da Figura 5.3 (a), um padrão de correlação semelhante é também observado entre as bases do grupo {A1, A2} em relação à base G1. Os valores de coeficiente de correlação entre as bases das três posições dos codões apresentados na Tabela 5.5 dão suporte às conclusões que obtivemos pela análise do biplot. Este comportamento das variáveis no biplot relaciona-se com o facto de que o aumento da frequência das bases em um dos grupos implica a diminuição das frequências das bases do grupo oposto (Lei de Chargaff).

Os biplots composicionais clássicos, representados nas Figuras 5.3 (b) e (c), referentes às frequências de bases nas três posições dos codões, também mostram padrões nas variáveis semelhantes aos observados nos biplots composicionais considerando apenas bases em cada uma das posições separadamente. Em particular, observamos dois grupos de variáveis, sendo um formado pelas bases do par (A, T) e outro pelas bases do par (C, G), em que as bases em cada grupo tendem a preservar ligações muito curtas entre si, mas ligações muito longas em relação às bases do grupo oposto. Por exemplo, as log-razões envolvendo as bases dos grupos {A1, A2, T1}, {A3, T3} e {C3, G3} são aproximadamente constantes e, pontanto, as bases pertencentes a cada um desses grupos são

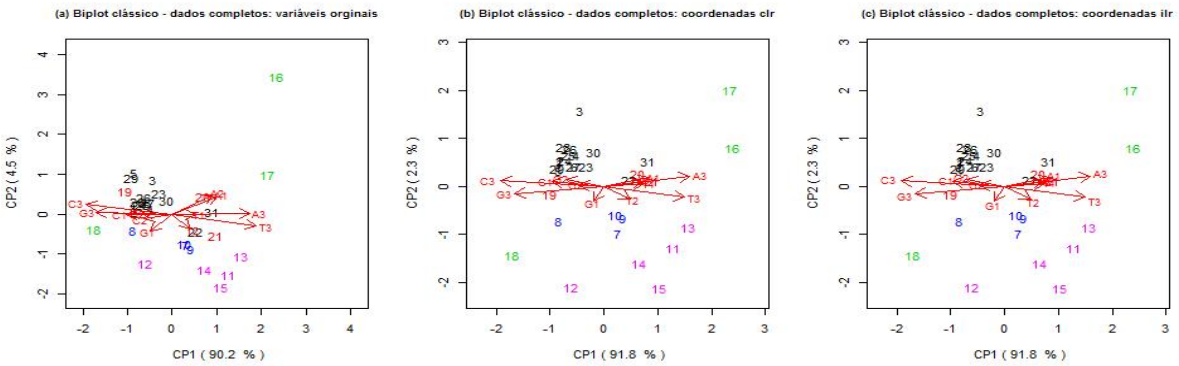


Figura 5.3. Biplots clássicos para as variáveis analisadas em termos absolutos (gráfico a) e composicionais (gráficos b e c) referentes às frequências das bases nas três posições dos codões.

Tabela 5.4. Valores dos desvios de frequências das bases nas três posições do codão, atestando o padrão de variação de bases das três posições do codão observados nos biplots para dados em coordenadas originais, apresentado na Figuras 5.3 (a), onde podemos ver que as bases da terceira posição apresentam valores de desvios mais elevados.

Bases	A1	C1	G1	T1	A2	C2	G2	T2	A3	C3	G3	T3
Variáveis	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}	x_{12}
Desvios padrão	0.046	0.045	0.031	0.025	0.045	0.030	0.031	0.025	0.075	0.081	0.072	0.080

Tabela 5.5. Tabela de correlações entre bases nas três posições do codão, reforçando as conclusões sobre o padrão de correlação entre bases obtidas pela análise do biplot clássico para variáveis originais representado na Figura5.3 (a).

Bases	A1	C1	G1	T1	A2	C2	G2	T2	A3	C3	G3	T3
A1	1.00	−0.90	−0.83	0.78	0.97	−0.90	−0.87	0.40	0.87	−0.85	−0.89	0.83
C1		1.00	0.59	−0.85	−0.87	0.89	0.95	−0.69	−0.91	0.95	0.95	−0.95
G1			1.00	−0.75	−0.74	0.62	0.56	−0.12	−0.68	0.60	0.68	−0.58
T1				1.00	0.67	−0.68	−0.810	0.62	0.85	−0.86	−0.88	0.86
A2					1.00	−0.90	−0.85	0.35	0.79	−0.78	−0.83	0.79
C2						1.00	0.83	−0.62	−0.83	0.87	0.87	−0.86
G2							1.00	−0.72	−0.84	0.88	0.91	−0.92
T2								1.00	0.62	−0.74	−0.68	0.77
A3									1.00	−0.96	−0.96	0.90
C3										1.00	0.97	−0.97
G3											1.00	−0.97
T3												1.00

proporcionais entre si. Além disso, o facto de as setas correspondentes a essas bases serem (aproximadamente) colineares no biplot composicional, um diagrama ternário envolvendo frequências das bases pertencentes a grupos opostos, conforme discriminados acima, exibirá um padrão de variação aproximadamente linear, conforme observamos na Figura 5.4.

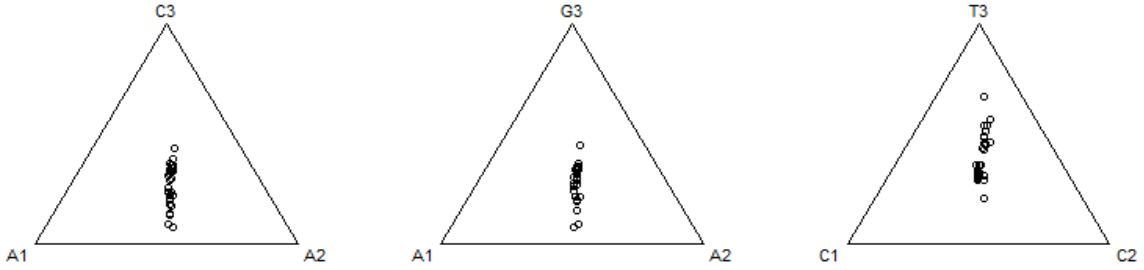


Figura 5.4. Diagramas ternários que mostram o padrão de variação linear exibidos pelas bases dos grupos {A1, A2, C3}, {A1, A2, G3}, e {C1, C2, T3}, cujas setas das bases pertencentes a cada grupo eram aproximadamente colineares nos biplots composicionais referentes às frequências de bases nas três posições do codão, apresentados nas Figuras 5.3 (b) e (c).

Um aspecto curioso nos biplots composicionais representados nas Figuras 5.3 (b) e (c), prende-se com o facto de os grupos formados pelas observações referentes à classe bactérias não ser tão coeso conforme observamos no biplot para dados originais. Isto pode ser consequência da presença de *outliers* no conjunto de dados, causando assim a distorção dos resultados. Por isso, construímos biplots robustos, para dados em coordenadas originais e em coordenadas ilr-transformadas, conforme representados na Figura 5.5, que são menos sensíveis à presença de *outliers* no conjunto de dados. Na figura da esquerda está o biplot robusto para dados originais, onde verificamos que os padrões das variáveis e das observações se mantêm inalteradas em comparação com o biplot clássico apresentado na Figura 5.3 (a). Isto permite-nos concluir que os resultados observados no biplot clássico para variáveis originais (Figura 5.3. (a)) não foram distorcidos pela presença de *outliers* no conjunto de dados e, portanto, são confiáveis para as análises subsequentes.

No biplot composicional robusto (Figura 5.5 à direita) verificamos alterações nos padrões das variáveis e das observações, em comparação com os biplots composicionais clássicos apresentados nas Figuras 5.3 (b) e (c). Por exemplo, no biplot composicional robusto verificamos a formação de três grupos de variáveis, cujas bases em cada uma estão fortemente correlacionadas entre si. Assim, temos o grupo {C1, G1, G2, G3} em oposição ao grupo {A1, A2, T1, T2, T3}, e o grupo {C2, C3} em oposição à base A3. Assim, as log-razões envolvendo bases dentro de cada um dos grupos acima discriminados tenderiam a apresentar pequenos valores de variância e, portanto, as frequências das bases envolvidas são aproximadamente proporcionais entre si, enquanto log-razões envolvendo bases de grupos opostos apresentariam maior variabilidade relativa entre si. Um diagrama ternário envolvendo bases pertencentes a grupos opostos, conforme discriminados acima, apresentaria um padrão de variação aproximadamente linear (ver Figura 5.6). Observamos também que o biplot composicional robusto separa as espécies em apenas dois grupos, sendo um grupo formado apenas pelas observações referentes às espécies do reino Animal, que privilegiam as bases C e G nos seus codões (espécies com valores negativos na CP1), e um grupo oposto constituído pelas observações referentes às classes das plantas, bactérias e fungos, que privilegiam as bases A e T nos seus codões (espécies com valores positivos de na CP1).

O biplot composicional robusto representado na Figura 5.5 (gráfico à direita), foi construído com recurso à função `mvoutlier.CoDa()`, disponível na biblioteca `mvoutlier` do **R** (Filzmoser *et al*, 2015). Além de construir biplots composicionais robustos (menos sensível à presença de *outliers* no conjunto dos dados), aquela função permite também a identificação dos *outliers*, representando-os

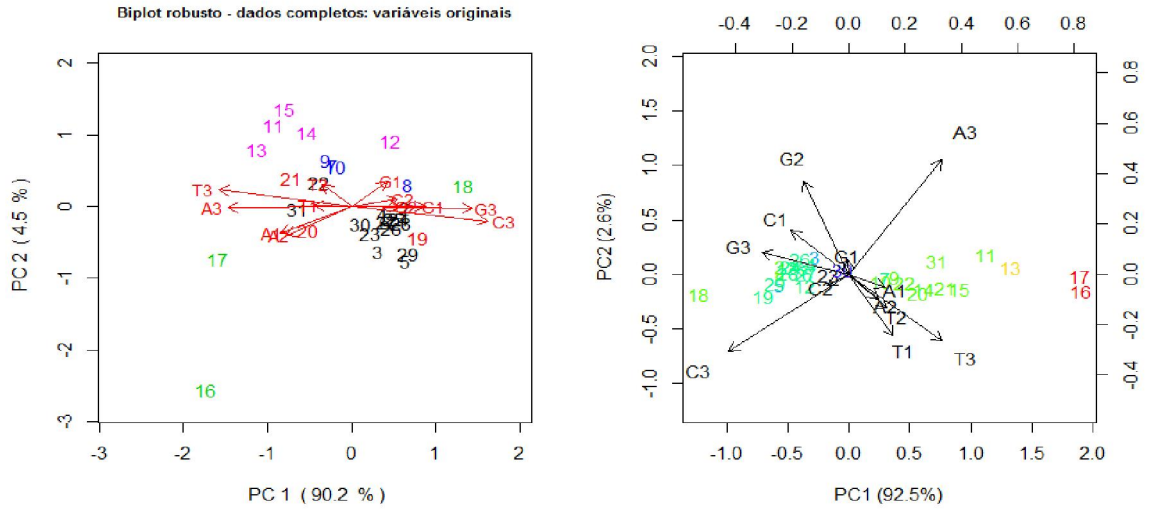
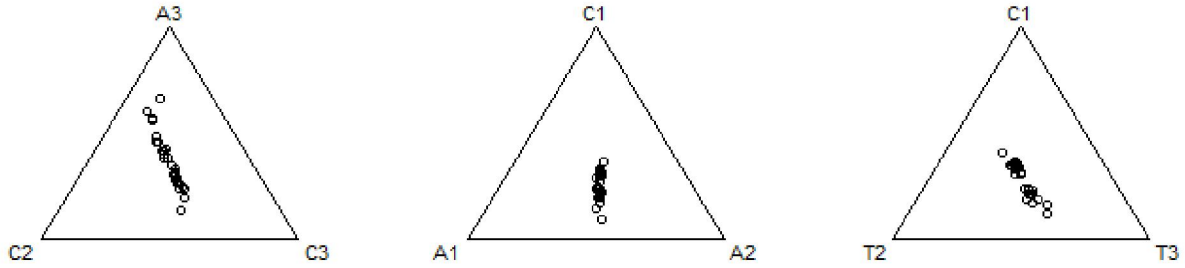


Figura 5.5. Biplot robusto referente às bases nas três posições dos codões, para dados em coordenadas originais (*esquerda*) e dados em coordenadas ilr –transformadas (*direita*), onde observamos padrões de variáveis e observações diferentes dos observados nos biplots composicionais clássicos, representados nas Figuras 5.3 (b) e (c).



(a) (b) (c)

Figura 5.6. Diagramas ternários que mostram o padrão de variação aproximadamente linear exibidos pelas bases dos grupos {C2, C3, A3}, {A1, C1, A2}, e {C1, T2, T3}, cujas setas das bases pertencentes a cada grupo estão dispostas de forma aproximadamente colineares no biplot composicional robusto referentes às frequências de bases das três posições do codão, representado na Figura 5.5 (*à direita*).

por cores progressivamente mais vivas (azul, verde, amarelo, vermelho), de acordo com a média das distâncias de Mahalanobis de cada observação em relação à média (origem dos eixos). Assim, observações com maiores valores da média das distâncias em relação ao centro dos dados são representadas pela cor vermelha, enquanto aquelas com valores mais baixos são representadas pela cor azul (Filzmoser et al, 2012). No caso em análise, os *outliers* identificados no conjunto de dados correspondem às observações referentes à Bactéria *Sa* (nº 13), e aos Protozoários *Pl* (nº 16) e *Dd* (nº 17).

Estudo 3: *Análise de dados fundidos – soma das frequências de cada uma das bases*

A fusão das frequências de cada uma das bases nas três posições de um codão permitiu a redução da dimensão dos dados para 4 componentes, identificadas pelas letras A, C, G e T, correspondentes às bases Adenina, Citosina, Guanina e Timina, respetivamente. Os biplots clássicos contruídos para os dados obtidos por esta fusão estão representados na Figura 5.7, em que o biplot da esquerda foi

aplicado aos dados em coordenadas originais, enquanto o biplot da direita foi aplicado aos dados em coordenadas clr-transformadas. No biplot para dados originais observamos a formação de apenas um grupo coeso, formado pelas espécies pertencentes ao reino Animal, com exceção das observações referentes aos animais *Eq* (nº3)*, Ce* (nº 22) e *Am* (nº 31), que se encontram dispersas no espaço 2-dimensional do biplot, assim como acontece com as espécies dos restantes reinos. Contudo, à semelhança do que observamos no Estudo 2, verificamos que as espécies do reino animal tendem a privilegiar as bases C e G, em oposição às bactérias, que tendem a privilegiar as bases A e T. Por outro lado, no biplot composicional, observamos a formação de dois grupos, sendo um formado pelos animais (preto) em oposição a outro formado pelas bactérias (magenta).

A percentagem de variabilidade retida pelo biplot para dados originais é de $95.2 + 3.6 = 98.8\%$, o que significa que a fidelidade na representação da estrutura dos dados pelo biplot é muito boa. Assim, os padrões das setas no biplot indicam-nos que, nos codões das espécies observadas, as bases A e C apresentam maiores valores de desvios padrão. O padrão de correlação entre as bases nos dados fundidos é semelhante ao que observámos nos Estudos 1 e 2. Neste caso, reportamos uma forte correlação positiva entre as bases C e G, enquanto estas, por sua vez, estão negativamente correlacionadas com as bases A e T.

No caso do biplot composicional (Figura 5.7, à esquerda), observamos também padrões semelhantes aos observados nos estudos anteriores, ou seja, temos dois grupos de partes, nomeadamente {A,T} e {C, G}, cujas ligações entre setas correspondentes às partes em grupos diferentes apresentam ligações muito longas entre si, enquanto as ligações entre as setas que pertencentes ao mesmo grupo consideravelmente mais curtas. Isto sugere que as log-razões envolvendo as partes de grupos opostos apresentariam maior variabilidade do que as log-razões envolvendo partes pertencentes ao mesmo grupo. E, como as ligações [A,C] e [T, G] são, aproximadamente, paralelas, significa que as log-razões $ln(A/C)$ e $ln(T/G)$ estão fortemente correlacionadas.

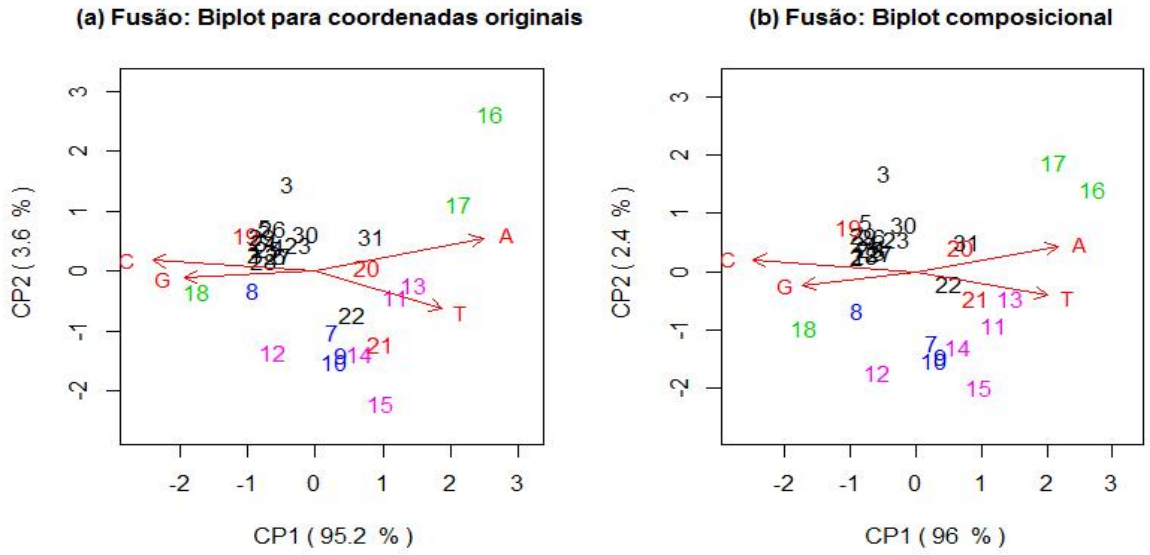


Figura 5.7. Biplot clássico para dados fundidos em coordenadas originais (*à esquerda*), e em coordenadas clr – transformadas.

Estudo 4: *Análise de dados fundidos – análise do teor C+G e A+T nas três posições dos codões*

A aplicação de biplots sobre dados fundidos pela soma do teor C+G e a A+T em cada uma das três posições dos codões das 31 espécies observadas destacou ainda mais o padrão de correlação entre os pares de nucleótidos (A, T) e (C, G), conforme podemos observar na Figura 5.8. O biplot da esquerda foi aplicado sobre dados fundidos em coordenadas originais, enquanto o biplot da direita foi aplicado sobre dados em coordenadas clr-transformadas.

No biplot aplicado sobre dados originais (Figura 5.8, à esquerda), observamos que os pontos exibem um padrão semelhante aos observados nos estudos precedentes, nomeadamente a separação das observações em dois grupos pela primeira componente principal, em termos do teor de CG em oposição ao teor de AT. Neste caso, verificamos que as espécies pertencentes ao reino Animal tendem a privilegiar as bases C e G nos seus codões (exceto as espécies 22 e 31), enquanto as bactérias (reino Monera) tendem a privilegiar as bases A e T (exceto a espécie 12).

A percentagem de variabilidade retida pelo biplot aplicado sobre dados originais é de $96.8 + 2.1 = 98.9\%$, o que significa que a fidelidade na representação da estrutura dos dados pelo biplot é muito boa (ver Figura 5.8, à direita). Os padrões das setas no biplot indicam-nos que, nos codões das espécies observadas, as bases na terceira posição são as que apresentam maiores valores de desvios padrão. O padrão de correlação entre as bases nos dados nesta fusão permite-nos observar que existe perfeita correlação negativa entre as frequências de AT e CG, em cada uma das três posições dos codões das 31 espécies observadas. Os coeficientes de correlações entre o teor AT e CG em cada uma das posições dos codões, representados na Tabela 5.6, apoiam estas conclusões.

No biplot composicional (Figura 5.8, à direita) observamos que as setas que representam as variáveis AT na primeira e na segunda posições apresentam ligações muito curtas entre si. O mesmo acontece com as setas que representam as variáveis CG. Isto significa que as frequências de cada um desses pares na primeira e na segunda posições dos codões têm log-razões aproximadamente contantes e, portanto, o teor de AT (ou CG) na primeira posição é aproximadamente proporcional ao teor de AT (respetivamente CG) na segunda posição dos codões das 31 espécies observadas.

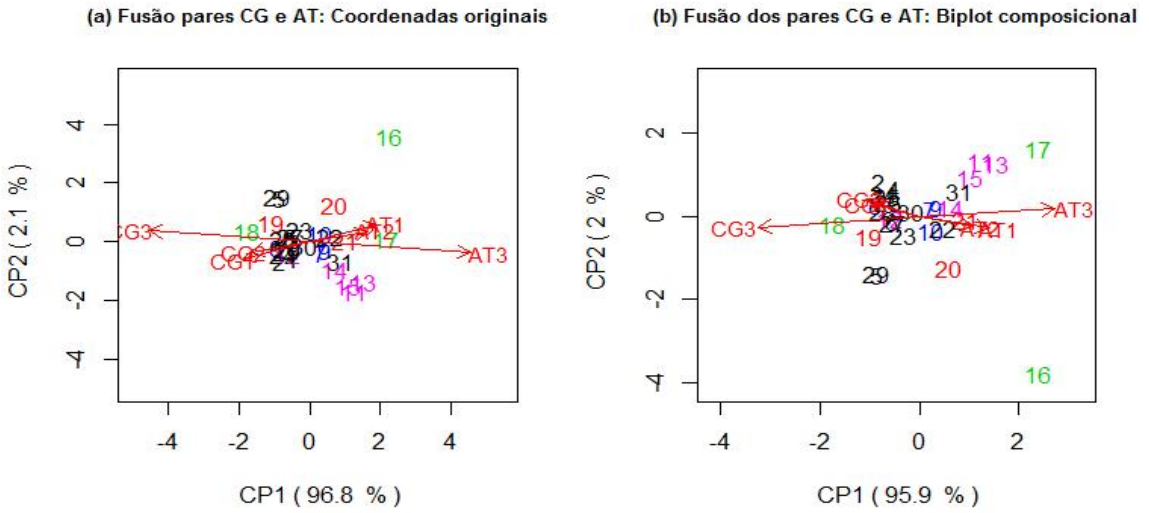


Figura 5.8. Biplot clássico para dados fundidos, em termos do teor de C+G e A+T, em coordenadas originais (*à esquerda*), e em coordenadas clr-transformadas.

Tabela 5.6. Tabela de correlações de dados fundidos pela soma A+T e C+G em cada uma das três posições dos codões, onde podemos observar perfeita correlação negativa entre as frequências de AT e CG em cada uma das três posições dos codões das 31 espécies observadas (ver Figura 5.8, *à esquerda*).

Bases	AT1	CG1	AT2	CG2	AT3	CG3
AT1	1.00	**−1.00**	0.91	−0.91	0.92	−0.92
CG1		1.00	−0.91	0.91	−0.92	0.92
AT2			1.00	**−1.00**	0.92	−0.92
CG2				1.00	−0.92	0.92
AT3					1.00	**−1.00**
CG3						1.00

As setas que representam os pares AT e CG da terceira posição apresentam ligações muito longas entre si, o que significa que existe grande variabilidade relativa entre as frequências dos pares AT e CG nesta posição dos codões das espécies observadas. A tabela de variação de log-razões (Tabela 5.7) reforça ainda mais essas conclusões.

Tabela 5.7. Tabela variação de log-razões referente aos dados fundidos pela soma A+T e C+G em cada uma das três posições dos codões, onde podemos observar que as log-razões entre os pares AT (e CG) na primeira e segunda posição apresentam valores variação muito pequenas (ver Figura 5.8, *à direita*).

Bases	AT1	CG1	AT2	CG2	AT3	CG3
AT1	−	0.078	**0.004**	0.083	0.029	0.281
CG1		−	0.055	**0.004**	0.174	0.076
AT2			−	0.066	0.041	0.243
CG2				−	0.188	0.071
AT3					−	**0.462**
CG3						−

A seguir, apresentamos quatro quadros resumos, contendo uma súmula de caraterísticas registadas em cada um dos quatro casos de estudos apresentados neste capítulo.

Quadro resumo das características mais relevantes observadas nos biplots relativos ao Estudo 1 (Figura 5.1):

	1ª Posição	2ª Posição	3ª Posição
	Biplot clássico – variação absoluta (dados em "bruto") – coordenadas originais		
Qualidade de representação	(86.9+9.3)%	(83.3 + 13.4)%	(96.6 + 2.5)%
Interpretação das CP's	CP1: contexto CG versus contexto AT CP2: conteúdo C versus conteúdo G; separa animais de bactérias.	CP1: contexto CG versus contexto AT CP2: conteúdo A versus conteúdo T; separa animais de bactérias.	CP1: contexto CG versus contexto AT CP2: conteúdo A versus conteúdo T;
Traço relevantes nas espécies	Bactérias com %G1>>	Bactérias com %T2>>	
Características dos nucleótidos	Maior dispersão sobre %A1 %A1 e %T1 fortemente correlacionados	Maior dispersão sobre %A2 %C2 e %G2 fortemente correlacionados	Maior dispersão sobre %C3 e %G3 %C3 e %G3 fortemente correlacionados
	Biplot clássico – variação relativa (natureza composicional dos dados) – coordenadas clr-transformadas		
Qualidade de representação	(90.1 + 6.8)%	(87 + 8.8)%	(97.2 + 1.6)%
Interpretação das CP's	CP1: contexto CG versus contexto AT CP2: conteúdo C versus conteúdo G; separa animais de bactérias.	CP1: contexto CG versus contexto AT CP2: conteúdo CT versus conteúdo AG; separa animais de bactérias.	CP1: contexto CG versus contexto AT CP2: conteúdo C versus conteúdo G; separa animais de bactérias.
Traço relevantes nas espécies	----	----	----
Características dos nucleótidos	%A1 ∝ %T1 %C1/%A1 fortemente correlacionada com %T1/%A1	---- %A2/G2 fortemente correlacionada com %T2/%C2	%C3 ∝ %G3 %C3/%A3 fortemente correlacionada com %G3/%A3

Quadro resumo das características mais relevantes observadas nos biplots relativos ao Estudo 2:

	Três posições dos nucleótidos no codão
	Biplot clássico – variação absoluta (dados em "bruto") – coordenadas originais
Qualidade de representação	(90.2+2.5)%
Interpretação das CP's	CP1: contexto CG versus contexto AT
	CP2: separa animais de bactérias.
Traços relevantes nas espécies	Bactérias com %T2>>; Plantas com %G1>>; protozoários suspeitos de serem observações atípicas
Características dos nucleótidos	Maior dispersão sobre %C3 e %T3
	%C3 e %T3, %A3 e %G3 e %A1 e %G1 fortemente correlacionados
	Biplot clássico – variação relativa (natureza composicional dos dados) – coordenadas clr
Qualidade de representação	(91.8 + 2.3)%
Interpretação das CP's	CP1: contexto CG versus contexto AT
	CP2: separa animais das bactérias e plantas.
Traços relevantes nas espécies	Bactérias e plantas com %T3/%A3>> em oposição com animais com %A3/%T3>>; protozoários suspeitos de serem observações atípicas
Características dos nucleótidos	%A1 $\propto$ %T1
	%G3/%G1, %G1/%T2, %T2/%T3, %C3/%A3 fortemente correlacionadas entre si.
	Biplot robusto – variação absoluta (dados em "bruto") – coordenadas originais
Qualidade de representação	(90.2+4.5)%
Interpretação das CP's	CP1: contexto CG versus contexto AT
	CP2: conteúdo C versus conteúdo G; separa animais de bactérias.
Traços relevantes nas espécies	Bactérias com %T2>>; Plantas com %G1>>; protozoários suspeitos de serem observações atípicas
Características dos nucleótidos	Maior dispersão sobre %C3 e %T3; Menor dispersão sobre %T1 e %T2.
	%G1 e %T2 não correlacionados
	%C3 e %T3, %A3 e %G3, %A1 e %G1, %A2 e %G1, %C1 e %T1 fortemente correlacionados
	Biplot robusto – variação relativa (natureza composicional dos dados) – coordenadas ilr
Qualidade de representação	(92.5 + 2.6)%
Interpretação das CP's	CP1: contexto CG versus contexto AT

	CP2: ---
Traços relevantes nas espécies	Protozoários suspeitos de serem observações atípicas
Características dos nucleótidos	%A1 ∝ %A2
	%C3/%G3,%C1/%G2; %A3/%A1, %A1/%A2 e %A3/%T1 fortemente correlacionadas entre si e não correlacionados com %G3/%A1

Quadro resumo das características mais relevantes observadas nos biplots relativos ao Estudo 3 (Figura 5.7):

	Fusão – soma das frequências de cada uma das bases
	Biplot clássico – variação absoluta (dados em "bruto") – coordenadas originais
Qualidade de representação	(95.2 + 3.6)%
Interpretação das CP's	CP1: contexto CG versus contexto AT
	CP2: separa animais das bactérias e plantas.
Traços relevantes nas espécies	Animais com %C e %G acima da média; protozoários suspeitos de serem observações atípicas
Características dos nucleótidos	Maior dispersão sobre %A
	%C e %G fortemente correlacionados
	Biplot clássico – variação relativa (natureza composicional dos dados) – coordenadas clr
Qualidade de representação	(96 + 2.4)%
Interpretação das CP's	CP1: contexto CG versus contexto AT
	CP2: separa animais das bactérias e plantas.
Traços relevantes nas espécies	Bactérias e plantas com %T/%A>>; protozoários suspeitos de serem observações atípicas
Características dos nucleótidos	%A/%C e %T%G altamente correlacionados

Quadro resumo das características mais relevantes observadas nos biplots relativos ao Estudo 4 (Figura 5.8):

	Fusão – análise em termos do teor C+G e A+T
	Biplot clássico – variação absoluta (dados em "bruto") – coordenadas originais
Qualidade de representação	(96.8 + 2.1)%
Interpretação das CP's	CP1: contexto CG versus contexto AT
	CP2: - -
Traços relevantes nas espécies	Animais com %CG>>; bactérias com %AT>>; Plantas com %G1>>; protozoários suspeitos de serem observações atípicas
Características dos nucleótidos	Maior dispersão sobre %AT3 e %CG3
	%AT e %CG fortemente correlacionados nas três posições dos codões
	Biplot clássico – variação relativa (natureza composicional dos dados) – coordenadas clr
Qualidade de representação	(95.9 + 2)%
Interpretação das CP's	CP1: contexto CG versus contexto AT
	CP2: - -
Traços relevantes nas espécies	Bactérias com %AT >> em oposição com animais com %CG>>; protozoários suspeitos de serem observações atípicas
Características dos nucleótidos	%AT1 $\propto$ %AT2 e %CG1 $\propto$ %CG2
	%CG1/%AT1 (ou %CG1/%AT2) fortemente correlacionado com %CG2/%AT1 (ou %CG2/%AT2)

Legenda:

- CP1: primeira componente principal
- CP2: segunda componente principal
- CP's: componentes principais
- $\propto$: proporcional a
- >>: acima da média

Conclusões e considerações finais

A análise estatística de dados composicionais é uma área relativamente recente e em desenvolvimento, que remonta aos anos 80, com os trabalhos de Aitchison (1986). Embora se tenha registado um crescimento de trabalhos e desenvolvimentos teóricos com o objetivo de propiciar a análise deste tipo de dados nas últimas duas décadas, a quantidade de trabalhos nessa área ainda é relativamente reduzida. A análise de dados com base em coordenadas log-razões transformadas ainda representa uma barreira à análise e interpretação de dados composicionais, não sendo muito divulgada, pelo que em muitos casos de estudos, a aplicação de técnicas multivariadas usuais, sem levar em conta a natureza composicional dos dados, continua a ser a opção adotada. No entanto, conforme mostramos neste trabalho, tais práticas podem levar a conclusões erradas, devido à singularidade da matriz de dados e presença de correlações espúrias.

Assim, aplicamos os biplots tradicional (para dados em bruto) e composicional para explorar informação absoluta e relativa contida num conjunto de dados do espaço dos codões, considerando-o, quer como dados multivariados reais sem restrições, quer levando em conta a sua natureza composicional. O conjunto de dados considerado contém as frequências relativas das quatro bases dos nucleótidos nas três posições dos codões de 31 espécies pertencentes aos cinco reinos de seres vivos, sendo: 16 animais, 4 plantas, 5 bactérias, 3 fungos e 3 protozoários. Nos quatro casos de estudo considerados, os biplots permitiram visualizar uma separação nítida entre as espécies pertencentes ao reino animal e as bactérias, sendo que os animais apresentam dominância dos nucleótidos C e G nos seus codões, enquanto que as bactérias apresentam dominância dos nucleótidos A e T. O biplot robusto, aplicado sobre dados originais e em coordenadas log-razões transformadas sugerem que os protozoários são observações atípicas no conjunto de dados considerados. Quanto às variáveis, a análise na perspetiva absoluta permitiu-nos observar que as bases da terceira posição dos codões são as que apresentam maiores valores de desvios padrão. Observamos também a existência de forte correlação positiva entre as bases A e T, e entre as bases C e G, enquanto que os pares (A, T) e (C, G) estão negativamente correlacionados. A análise na perspetiva relativa permitiu-nos concluir que, no caso das 31 espécies consideradas, as frequências das bases A e T são, aproximadamente, proporcionais entre si, verificando-se o mesmo padrão de variabilidade relativa em relação às frequências das bases C e G. No entanto, existe uma grande variação relativa entre as bases do par (A, T) em relação às bases do par (C, G).

A aplicação de biplots aos dados em coordenadas ilr-transformadas exige a construção de uma base ortonormal no simplex, constituída por um conjunto de vetores ortonormais de cardinalidade igual à caraterística da matriz dos dados. Para os biplots composicionais clássicos utilizamos bases ortonormais determinadas pela partição binária sequencial (PBS) segundo Egozcue *et al* (2005), enquanto que para o biplot composicional robusto utilizamos uma base ortonormal determinada pela PBS segundo Filzmoser *et al* (2009), implementadas, respetivamente, nos pacotes `Compositions` e `mvoutlier` do software **R**. No entanto, no caso do espaço dos codões, onde cada vetor contém 12 componentes, que satisfazem a condição

$$x_1 + x_2 + x_3 + x_4 = x_5 + x_6 + x_7 + x_8 = x_9 + x_{10} + x_{11} + x_{12} = h,$$

em que h representa o número total dos codões no genoma, variável de espécie para espécie, a caraterística da matriz de dados é 9, visto que x_4 pode ser escrito como combinação linear de x_1, x_2 e x_3; x_8 pode ser escrito como combinação linear de x_5, x_6 e x_7; e x_{12} pode ser escrito como combinação

linear de x_9, x_{10} e x_{11}. Dada esta redundância peculiar dos vetores que constituem o espaço dos codões, a determinação de uma base ortonormal (constituída por 11 vetores) seguindo o processo de PBS estabelecido na literatura atual, descrito na Seção 2.3.5 e implementados nos referidos pacotes do **R**, não se perspetiva adequado na construção dos biplots robustos para os dados do espaço dos codões em coordenadas ilr-transformadas. A base deveria ser constituída por 9 vetores e não por 11! Assim, a transformação ilr dos dados a partir de uma base ortonormal e a construção de biplots composicionais robustos carecem de uma investigação mais aprofundada em situações semelhantes ao do espaço dos codões. Na realidade, o espaço dos codões constitui um caso particular de dados composicionais, ao qual poderemos dizer que corresponde a uma mistura de composições onde cada composição entra com igual peso ou também a uma composição de composições. Na literatura não encontramos nenhuma referência nem procedimento específico de análise para este tipo de dados. Assim, no futuro pretendemos desenvolver técnicas adequadas para determinação de bases ortonormais e análise de conjuntos de dados composicionais com aquela caraterística, em particular ao espaço dos codões e outras composições de composições.

Referências

Aitchison, J. (1986) *The Statistical Analysis of Compositional Data.* London: Chapman and Hall.

Aitchison, J., Greenacre, M. (2002) Biplots of compositional data. *Appl. Statist.,* 51(4), 375-392.

Aitchison, J. (2005) *A concise Guide to Compositional Data Analysis.* 2nd Compositional Data Analysis Workshop – CoDaWork'05. Disponível em: http://www.leg.ufpr.br/lib/exe/fetch.php/pessoais:abtmartins:a_concise_guide_to_compositional_data_analysis.pdf

Buccianti, A., Mateu-Figueras, G., and Pawlowsky-Glahn, V. (2006) *Compositional Data Analysis in the Geosciences: From Theory to Practice*, Special Publications, vol. 264, Geological Society, London, 212 p.

Egozcue, J.J., Pawlowsky-Glahn, V. (2005) Groups of parts and their balances in compositional data analysis. *Mathematical Geology*, 37(7), 795–828.

Filho, D., Júnior, J. (2009) Desvendando o Mistério do Coeficiente de Correlação de Pearson. *Revista Política Hoje*, 18(1), 115-146.

Filzmoser, P., Hron, K., Reimann, C. (2009) Principal component analysis for compositional data with outliers. *Environmetrics* 20 (6), 621–632.

Filzmoser, P., Hron, K., Reimann, C. (2012) Interpretation of multivariate outliers for compositional data. *Computer & Goesciences* 39, 77-85.

Filzmoser, P., Gschwandtner, M. (2015). mvoutlier: Multivariate outlier detection based on robust methods. R package version 2.0.6. http://CRAN.R-project.org/package=mvoutlier

Gabriel, K. (1971) The biplot grephif display of matrices with application to principal components analysis. *Biometrika* 58(3), 453-467.

Gallo, M. (2007) The Scaling Problems in Service Quality Evaluation. *Metod. Zvezki* 4(2), 165-176

Greenacre, M. Principal Components Analysis Biplot. In: Greenacre, M. (2010) *Biplot in Practice*. Disponível em: http://www.fbbva.es/TLFU/dat/greenacre_c06_2010.pdf

Hron, k., Jelínková, M., Filzmoser, P., Kreuziger, R., Bednár, P., Barták, P. (2012) Statistical analysis of wines using a robust compositional biplot. *Talanta* 90, 46-50.

Hron, K. (2012) *Classical and robust statistical methods for a comprehensive statistical treatment of compositional data*. Habilitation Thesis. Disponível em: https://is.muni.cz/do/rect/habilitace/1431/Hron/habilitace/habilitation_thesis-Hron.pdf?lang=en

Insana, G. (2003) *DNA Phonology: Investigating the Codon Space.* Tese de doutoramento. Disponível em: https://www.ebi.ac.uk/sites/ebi.ac.uk/files/shared/documents/phdtheses/giuseppeinsanathesis.pdf

Kynclová, P., Filzmoser, P., Hron, K. (2015) Compositional biplots including external non-compositional variables. *Statistics*, 1-18.

Kohler, U., Luniak, M. (2005) Data inspection using biplots. *The Stata Journal* 5(2), 208-223.

Kucera, M., Malmgren, B. (1998) Logratio transformation of compositional data — a resolution of the constant sum constraint. *Marine Microp*. 34, 117-120.

Maronna, R., Martin, R., Yohai, V. (2006) *Robust Statistics: Theory and methods.* John Wiley & Son

Pawlowsky-Glahn, V., Egozcue, J.J. (2006) Compositional Data and Their Analysis: An Introduction, Special Publications, *Geological Society of London, Special Publication*, 264, 1-10.

Pawlowsky-Glahn, V., Buccianti, A. (2011) *Compositional Data Analysis: Theory and Applications*. John Wiley & Sons.

Pawlowsky-Glahn, V., Egozcue, J.J., Tolosana-Delgado, R. (2015) *Modeling and Analysis of Compositional Data.* John Wiley & Sons.

R Core Team (2014). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/

Ripley, B., Lapsley, M. (2014). RODBC: ODBC Database Access. R package version 1.3-10. http://CRAN.R-project.org/package=RODBC

Rousseeuw, P., Driessen, K. (1999) A Fast algorithm fot the Minimum Covariance Determinant Estimator. *Technometrics*, 41(3), 212-223.

Takeuchi, F., Futamura, Y., Yoshikura, H., Yamamoto, K. (2003) Statistics of trinucleotides in coding sequences and evolution. *Journal of Theor. Bio*. 222, 139-149.

Todorov, V., Filzmoser, P. (2009). An Object-Oriented Framework for Robust Multivariate Analysis. *Journal of Statistical Software*, 32(3), 1-47. URL http://www.jstatsoft.org/v32/i03/

van den Boogaart, K.G., Tolosana-Delgado, R. (2013) *Analysing Compositional Data with R*, Springer, Heidelberg.

van den Boogaart, K. G., Raimon Tolosana-Delgado, R., Bren, M. (2014). compositions: Compositional Data Analysis. R package version 1.40-1. http://CRAN.R-project.org/package=compositions

Wedlake, R. (2008) *Robust Principal Component Analysis Biplots.* Tese de Mestrado. Disponível em http://scholar.sun.ac.za/handle/10019.1/2491

Weir, B. (1996) *Genetic Data Analysis II: Methods for Discret Population Genetic Data.* Sinauer Associates, Inc. Publishers. Sunderland, Massachusetts.

Anexos

A.1. Lista das 31 espécies consideradas

Tabela A.1. Lista das 31 espécies consideradas, com a designação abreviada de cada espécie e indicação do seu domínio.

Bt	animal	Bos taurus (vaca)
Cf	animal	Cannis familiaris (cão)
Eq	animal	Equus caballus (cavalo)
Gg	animal	Gallus gallus (galinha)
Dm	animal	Drosophila melanogaster (mosca da fruta)
Um	animal	Ursus maritimus (urso-polar)
At	plant	Arabidopsis thaliana
Os	plant	Oryza sativa
Po	plant	Populus trichocarpa
Vv	plant	Vitis vinifera
Ba	bacteria	Bacillus anthracis Ames
Ec	bacteria	E. coli
Sa	bacteria	Staphylococcus aureus
St	bacteria	Streptococcus pneumoniae
Sm	bacteria	Streptococcus mutans
Pl	protozoa	Plasmodium falciparum (protozoário)
Dd	protozoa	Dictyostelium discoideum (protozoário)
Lm	protozoa	Leishmania major (protozoário)
Nc	fungi	Neurospora crassa (fungo)
Sc	fungi	Saccharomyces cerevisiae (fungo)
Sp	fungi	Schizosaccharomyces pombe OLD (fungo)
Ce	animal	Caenorhabditis elegans (minhoca)
Dr	animal	D rerio (peixe)
Hs	animal	H sapiens (primata)
Mm	animal	Macaca mulatta (primata)
Pt	animal	Pan troglodytes (primata)
Rn	animal	Rattus norvegicus (rato)
Ao	animal	Aotus nancymaae (macaco)
Fu	animal	Takifugu rubripes (peixe)
Xt	animal	Xenopus Tropicalis (sapo)
Am	animal	Apis mellifera (abelha)

A.2. Frequências absolutas das bases

Tabela A.2. Frequências absolutas dos quatro nucleótidos em cada uma das três posições dos codões das 31 espécies consideradas.

Abrev	A1	C1	G1	T1	A2	C2	G2	T2	A3	C3	G3	T3	classe
Bt	4658584	4735637	5824063	3076274	5275784	4627886	3872502	4518386	3552650	5514026	5335366	3892516	animal
Cf	4270226	4444220	5415317	2819113	4801452	4346744	3666201	4134479	3309939	5076311	4914293	3648333	animal
Eq	8883711	7269891	9082593	5067918	9278063	7308997	6160469	7556584	7529641	8251562	7883649	6639261	animal
Gg	4077037	3697159	4691313	2613071	4486660	3727900	3121716	3742304	3387264	3995958	4101493	3593865	animal
Dm	4685667	4211198	5286714	2893923	5471298	4080685	3173575	4351944	3007638	5389505	5202831	3477528	animal
Um	4742627	4208008	5350453	3047808	5276496	4176474	3403064	4492862	3665696	4854639	4731767	4096794	animal
Ay	3044649	1941081	3335369	2049869	3367848	2306139	1804456	2892525	2632122	1986146	2359476	3393224	Planta
Os	4717870	4362946	6783744	3208727	5511522	4861048	3885512	4815205	3358383	5800887	5765052	4148965	Planta
Po	954722	604222	970937	680836	1002662	702880	576896	928279	853062	592420	688148	1077087	Planta
Vv	579533	384655	587422	427014	605284	427924	356622	588794	502691	402104	446573	627256	Planta
Ba	400755	197102	422072	259564	435954	250347	183976	409216	488800	135819	195103	459771	bactéria
Ec	384287	373692	541765	240155	449796	348838	277428	463837	276289	418464	445615	399531	bactéria
Sa	236173	100274	223404	148261	256345	136365	93443	221959	273795	73386	83601	277330	bactéria
St	172728	103680	198278	113579	200900	119761	85091	182513	165342	105758	101400	215765	bactéria
Sm	172384	94703	180580	109272	192156	113024	77555	174204	156229	75969	84488	240253	bactéria
Pl	1889686	393912	912957	911597	2084188	492383	418931	1112650	1592311	321926	388104	1805811	Protozoário
Dd	2734929	917146	1579447	1702481	2812984	1399875	824840	1896304	2913480	534614	441576	3044333	Protozoário
Lm	1111886	1407036	1912587	808325	1367004	1544653	1050378	1277799	553523	1924545	1984897	776869	Protozoário
Nc	1022446	943722	1337037	631616	1196547	1060636	733570	944068	601583	1401671	1124222	807345	Fungo
SC	999089	487251	848733	672111	1058953	672761	430456	845014	883702	580863	558741	983878	Fungo
Sp	25330	14846	21642	19708	27564	17756	11815	24391	25367	13046	12755	30358	fungo
Ce	783112	503554	713358	566796	810887	590039	395655	770239	753690	492031	496646	824453	animal
Dr	4918570	3912104	5143233	3021236	5476245	3923830	3165298	4429770	3832009	4399449	4425226	4338459	animal
Hs	4786100	4670846	5892350	3057483	5456568	4593833	3860580	4495798	3739293	5342154	5283325	4042007	animal
Mm	1616792	1539715	1908759	1060980	1798896	1521860	1288579	1516911	1282546	1751231	1711184	1381285	animal
Pt	4568482	4072942	5412739	2835732	5130228	3998986	3663588	4097093	3685467	4628923	4879639	3695866	animal

Rn	4466047	3981873	5049055	2874750	4957896	3928536	3230949	4254344	3435921	4621089	4482906	3831809	animal
Ao	4015450	3956509	4680787	2781380	4400117	3851635	3255522	3926852	3068625	4787383	4236962	3341156	animal
Fu	3630062	3234455	4172166	2267321	4120254	3136229	2580432	3467089	2341721	4325518	4017106	2619659	animal
Xt	4708964	3689571	4898339	2927672	5162455	3855239	2996741	4210111	4067169	3984445	3925172	4247760	animal
Am	1306909	717307	1167165	836910	1408294	874630	656625	1088742	1369230	591177	725047	1342837	animal

A.3. Script em R

```
################################################################################
#              Análise Estatística de Dados Composicionais
#                     Mestrado em Matemática e Aplicações
#                                UA
#                             2015-2016
#                            Rodney Sousa
################################################################################
#      Exemplo 2.1. Correlação espúria
#      Tabela 2.1: Dados adaptados de Aitchison,2005
#             A: amostra observada pelo cientista A
#             B: amostra observada pelo cientista B
#-------------------------------------------------------------------------------
A=matrix(data=c(0.1,0.2,0.3,0.2,0.1,0.3,0.1,0.1,0.2,0.6,0.6,0.2),nrow=3,ncol=4)
B=matrix(data=c(0.25,0.4,0.43,0.5,0.20,0.43,0.25,0.40,0.14),nrow=3,ncol=3)
#-------------------------------------------------------------------------------
#      Tabela 2.3: Matriz de covariâcias para A e para B
#-------------------------------------------------------------------------------
cov.A=cov(A)        # matrix de covariância A
cov.B=cov(B)        # matrix de covariância B
#-------------------------------------------------------------------------------
#      Soma de cov(xi,xj), i<>j em cada linha de cov(A)
#-------------------------------------------------------------------------------
sum(cov.A[1,-1])
sum(cov.A[2,-2])
sum(cov.A[3,-3])
sum(cov.A[4,-4])
#-------------------------------------------------------------------------------
#      Soma de cov(xi,xj), com i<>j, em cada linha de cov(B)
#-------------------------------------------------------------------------------
sum(cov.B[1,-1])
sum(cov.B[2,-2])
sum(cov.B[2,-3])
#-------------------------------------------------------------------------------
#      Tabela 2.4: Matriz de correlações para A e para B
#-------------------------------------------------------------------------------
cor(A,method='pearson')          # Matriz de correlação A
cor(B,method='pearson')          # matriz e correlação B

################################################################################
#      Tabela 2.9: Coordenadas ilr-transformadas para A e B
#-------------------------------------------------------------------------------
ilr.A=matrix(0,3,3)     # Matrix de coordenadas ilr para A
for(i in 1:3){
       ilr.A[i,1]=(1/2)*log((A[i,1]*A[i,2])/(A[i,3]*A[i,4]))
       ilr.A[i,2]=(1/sqrt(2))*log(A[i,1]/A[i,2])
       ilr.A[i,3]=(1/sqrt(2))*log(A[i,3]/A[i,4])
}
ilr.A                      # Coordenadas ilr para A
#-------------------------------------------------------------------------------
ilr.B=matrix(0,3,2)     # Matrix de coordenadas ilr para B
for(i in 1:3){
       ilr.B[i,1]=log((B[i,1]*B[i,2])^(1/sqrt(6))/(B[i,3]^sqrt(2/3)))
       ilr.B[i,2]=(1/sqrt(2))*log(B[i,1]/B[i,2])
}
ilr.B                      # Coordenadas ilr para B
################################################################################
#      Exemplo 4.1: Centro da amostra registada pelo cientista A
#
```

```
N=3                         # N° de amostras
cen.gm=numeric(4)           # Vetor com 4 componentes
for(i in 1:4){
      cen.gm[i]=(prod(A[,i]))^(1/3)
}
cen.gm
cen=cen.gm/sum(cen.gm)
cen
#############################################################################
#     Exemplo 4.2: Tabela de variação referente aos dados do Cientista A
#*****************************************************************************
tv=matrix(0,4,4)                  # Tabela de variação
      #***** triângulo inferior: médias  ******
tv[2,1]=mean(log(A[,1]/A[,2]));tv[3,1]=mean(log(A[,1]/A[,3]));tv[4,1]=mean(log(A[,1
]/A[,4]))
tv[3,2]=mean(log(A[,2]/A[,3]));tv[4,2]=mean(log(A[,2]/A[,4]))
tv[4,3]=mean(log(A[,3]/A[,4]))

      #****** Triângulo superior: variâncias  ********
tv[1,2]=var(log(A[,1]/A[,2]));tv[1,3]=var(log(A[,1]/A[,3]));tv[1,4]=var(log(A[,1]/A
[,4]))
tv[2,3]=var(log(A[,2]/A[,3]));tv[2,4]=var(log(A[,2]/A[,4]))
tv[3,4]=var(log(A[,3]/A[,4]))

tv                          # Tabela 4.2
#############################################################################
#            Análise Estaística de Dados Composicionais
#            Capítulo 5: Aplicação ao Espaço dos Codões
#
#############################################################################
# ***** Bibliotecas com funções algumas funções Necessárias ******

require(RODBC)                          # Package conexão com o Excel
library(compositions)
library(rrcov)
library(mvoutlier)
#library(robCompositions)
#----------------------------------------------------------------------------
#            LEITURA DOS DADOS DO EXCEL
#----------------------------------------------------------------------------
ficheiro=odbcConnectExcel("CodonSpaceVersionSPE.xls")
dados=sqlFetch(ficheiro, sqtable="FreqRelSoma3")
x=dados[,2:13]; # head(x)             # Tabela A.2
cor=dados[,15]              # 1: preto, 2: vermelho, 3: verde, 4: azul, 6: Magenta
#============================================================================
#     ESTUDO 1: Frequências relativas das bases em cada posição de um codão,
#               de forma separada
#============================================================================
#----------------------------------------------------------------------------
#     CONSTRUÇÃO DE BIPLOT CLÁSSICOS
#----------------------------------------------------------------------------
#     VARIÁVEIS:
#      - invd: matriz de entradas 1/sd, em q sd=desvios
#      - load: loadings, dado por rotations*sd
#      - x.x: matriz dos scores, dada por xx=scores.pca*invd
#      - b: valores próprios da matriz de covariâncias
#----------------------------------------------------------------------------
#     1.1. Figura 5.1. FREQUÊNCIAS DE BASES NA 1ª POSIÇÃO
#----------------------------------------------------------------------------
      #*************************************************#
      #     Figura 5.1.(a) DADOS BRUTOS (VAR. ORIGINAIS)    #
```

```
      #*****************************************************#
      par(mfrow=c(1,3))
x1=dados[,2:5]      # x2=dados[,6:9];x3=dados[,10:13]; Bases na 1ª posição
   x1=scale(x1, center=TRUE, scale=FALSE)
   s1=svd(x1);pc.x1=prcomp(x1,retx=TRUE,center=TRUE);b.x1=summary(pc.x1) #SVD e PCA
      U1=s1$u; V1=s1$v; D1=diag(s1$d)
G1=sqrt(30)*U1                                          # X=GH', com G=U e H=VD
H1=sqrt(30)*V1%*%D1*1.6
      rownames(H1,do.NULL=TRUE,prefix="col")  # Nomes das variáveis
      rownames(H1)=colnames(x1)

#            REPRESENTAÇÃO DOS PONTOS
plot(G1[,1],G1[,2], main="Bases na 1ª posição: variáveis originais", cex.main=1,
     xlim=c(min(G1[,1],H1[,1],G1[,2],H1[,2],0)*1.1,max(G1[,1],H1[,1],G1[,2],H1[,2],0)*1.2),
      ylim=c(min(G1[,1],H1[,1],G1[,2],H1[,2],0)*1.1,max(G1[,1],H1[,1],G1[,2],H1[,2],0)*1.2),
     xlab=paste("(a) CP1 (",(round(100*b.x1$importance[2,1],digits=1)), " % )"),
     ylab=paste(" CP2 (", (round(100*b.x1$importance[2,2],digits=1)), " % )"), type="n")
     text(G1[,1],G1[,2],lab=rownames(x),col=cor)
      #     REPRESENTAÇÃO DAS SETAS
for(i in 1:4){
      arrows(0,0,H1[i,1],H1[i,2],col="red", length=0.1)
}
text(H1[,1]*1.15,H1[,2]*1.1,lab=colnames(x1),cex=0.9,col="red")

      #************************************************************#
      #     Figura 5.1.(b) DADOS EM COORDENADAS CLR-TRANSFORMADAS #
      #************************************************************#
x1.clr=clr(dados[,2:5]); x1.clr=data.frame(x1.clr)    # Transformação clr de x1
      y1=scale(x1.clr, center=TRUE, scale=FALSE)
      s1.clr=svd(y1); pc.x1.clr=prcomp(y1); b.x1.clr=summary(pc.x1.clr)
G1=pc.x1.clr$x%*%diag(sqrt(31-1)/s1.clr$d)
H1=pc.x1.clr$rotation%*%diag(s1.clr$d)*1.3

      #     REPRESENTAÇÃO DOS PONTOS
plot(G1[,1],G1[,2], main="Bases na 1ª posição: coordenadas clr", cex.main=1,
      xlim=c(min(G1[,1],H1[,1],G1[,2],H1[,2],0)*1.1,max(G1[,1],H1[,1],G1[,2],H1[,2],0)*1.2),
      ylim=c(min(G1[,1],H1[,1],G1[,2],H1[,2],0)*1.1,max(G1[,1],H1[,1],G1[,2],H1[,2],0)*1.2),
      xlab=paste("(b) CP1 (",(round(100*b.x1.clr$importance[2,1],digits=1)), " % )"),
      ylab=paste(" CP2 (", (round(100*b.x1.clr$importance[2,2],digits=1)), " % )"),type="n")
      text(G1[,1],G1[,2],lab=rownames(x),col=cor)
      #     REPRESENTAÇÃO DAS SETAS
for(i in 1:4){
      arrows(0,0,H1[i,1],H1[i,2],col="red", length=0.1)
}
text(H1[,1]*1.15,H1[,2]*1.1,lab=colnames(x1),cex=0.9,col="red")

      #************************************************************#
      #     Figura 5.1.(c) DADOS EM COORDENADAS ILR-TRANSFORMADAS      #
      #************************************************************#
x1.ilr=ilr(dados[,2:5]); z1=data.frame(x1.ilr)        # Transformação clr de x1
z1=scale(z1, center=TRUE, scale=FALSE); phi1=ilrBase(dados[,2:5],x1.ilr,4)
s1.z1=svd(z1); pc.x1.ilr=prcomp(z1); b.x1.ilr=summary(pc.x1.ilr)
loadz1=phi1%*%pc.x1.ilr$rotation; loadz1=cbind(loadz1,rep(0,4))   #
zzr=pc.x1.ilr$x%*%t(phi1)

G1=sqrt(30)*s1.z1$u # %*%diag(s1.clr$d)
H1=loadz1%*%diag(s1.clr$d)*1.4
plot(G1[,1],G1[,2], main="Bases na 1ª posição: coordenadas ilr", cex.main=1,
      xlim=c(min(G1[,1],H1[,1],G1[,2],H1[,2],0)*1.1,max(G1[,1],H1[,1],G1[,2],H1[,2],0)*1.2),
      ylim=c(min(G1[,1],H1[,1],G1[,2],H1[,2],0)*1.1,max(G1[,1],H1[,1],G1[,2],H1[,2],0)*1.2),
      xlab=paste("(c) CP1 (",(round(100*b.x1.ilr$importance[2,1],digits=1)), " % )"),
```

```
       ylab=paste(" CP2 (", (round(100*b.x1.ilr$importance[2,2],digits=1)), " % )"),type="n")
       text(G1[,1],G1[,2],lab=rownames(x),col=cor)
       #      REPRESENTAÇÃO DAS SETAS
for(i in 1:4){
       arrows(0,0,H1[i,1],H1[i,2],col="red", length=0.1)
}
text(H1[,1]*1.3,H1[,2]*1.3,lab=colnames(x1),cex=0.9,col="red")

#=============================================================================
#      1.2. FREQUÊNCIAS DE BASES NA 2ª POSIÇÃO
#=============================================================================
       #*********************************************************#
       #      Figura 5.1.(d) DADOS BRUTOS (VAR. ORIGINAIS)       #
       #*********************************************************#
       par(mfrow=c(1,3))
x2=dados[,6:9]      # x2=dados[,6:9];x3=dados[,10:13]; Bases na 1ª posição
       x2=scale(x2, center=TRUE, scale=FALSE)
       s2=svd(x2);pc.x2=prcomp(x2,retx=TRUE,center=TRUE);b.x2=summary(pc.x2) # SVD e PCA
       U2=s2$u; V2=s2$v; D2=diag(s2$d)
G2=sqrt(30)*U2                                        #      X=GH', com G=U e H=VD
H2=sqrt(30)*V2%*%D2*1.5                 # multiplicação por uma constante de escala
#      rownames(H2,do.NULL=TRUE,prefix="col")  # Nomes das variáveis
#      rownames(H2)=colnames(x2)

       #      REPRESENTAÇÃO DOS PONTOS
plot(G2[,1],G2[,2], main="Bases na 2ª posição: variáveis orginais", cex.main=1,
       xlim=c(min(G2[,1],H2[,1],G2[,2],H2[,2],0)*1.1,max(G2[,1],H2[,1],G2[,2],H2[,2],0)*1.2),
       ylim=c(min(G2[,1],H2[,1],G2[,2],H2[,2],0)*1.1,max(G2[,1],H2[,1],G2[,2],H2[,2],0)*1.2),
       xlab=paste("(d) CP1 (",(round(100*b.x2$importance[2,1],digits=1)), " % )"),
       ylab=paste(" CP2 (", (round(100*b.x2$importance[2,2],digits=1)), " % )"), type="n")
       text(G2[,1],G2[,2],lab=rownames(x),col=cor)
       #      REPRESENTAÇÃO DAS SETAS
for(i in 1:4){
       arrows(0,0,H2[i,1],H2[i,2],col="red", length=0.1)
}
text(H2[,1]*1.22,H2[,2]*1.1,lab=colnames(x2),cex=0.9,col="red")

       #*********************************************************#
       # Figura 5.1.(e) DADOS EM COORDENADAS CLR-TRANSFORMADAS    #
       #*********************************************************#
x2.clr=clr(dados[,6:9]); x2.clr=data.frame(x2.clr)    # Transformação ilr de x1
y2=scale(x2.clr, center=TRUE, scale=FALSE)
s2.clr=svd(y2); pc.x2.clr=prcomp(y2); b.x2.clr=summary(pc.x2.clr)
G2=pc.x2.clr$x%*%diag(sqrt(31-1)/s2.clr$d)
H2=pc.x2.clr$rotation%*%diag(s2.clr$d)*2.5    # constante d escala = 2.5
plot(G2[,1],G2[,2], main="Bases na 2ª posição: coordenadas clr", cex.main=1,
       xlim=c(min(G2[,1],H2[,1],G2[,2],H2[,2],0)*1.1,max(G2[,1],H2[,1],G2[,2],H2[,2],0)*1.2),
       ylim=c(min(G2[,1],H2[,1],G2[,2],H2[,2],0)*1.1,max(G2[,1],H2[,1],G2[,2],H2[,2],0)*1.2),
       xlab=paste("(e) CP1 (",(round(100*b.x2.clr$importance[2,1],digits=1)), " % )"),
       ylab=paste(" CP2 (", (round(100*b.x2.clr$importance[2,2],digits=1)), " % )"),type="n")
       text(G2[,1],G2[,2],lab=rownames(x),col=cor)
       #      REPRESENTAÇÃO DAS SETAS
for(i in 1:4){
       arrows(0,0,H2[i,1],H2[i,2],col="red", length=0.1)
}
text(H2[,1]*1.22,H2[,2]*1.1,lab=colnames(x2),cex=0.9,col="red")

       #*********************************************************#
       # Figura 5.1.(f) DADOS EM COORDENADAS ILR-TRANSFORMADAS    #
       #*********************************************************#
x2.ilr=ilr(dados[,6:9]); z2=data.frame(x2.ilr)        # Transformação ilr de x2
```

```
z2=scale(z2, center=TRUE, scale=FALSE); phi2=ilrBase(dados[,6:9],x2.ilr,4)
s2.z2=svd(z2); pc.x2.ilr=prcomp(z2); b.x2.ilr=summary(pc.x2.ilr)
loadz2=phi2%*%pc.x2.ilr$rotation; loadz2=cbind(loadz2,rep(0,4))
       #zzr=pc.x2.ilr$x%*%t(phi2)

G2=sqrt(30)*s2.z2$u # %*%diag(s1.clr$d)
H2=loadz2%*%diag(s2.clr$d)*1.5 #*1.4
plot(G2[,1],G2[,2], main="Bases na 2ª posição: coordenadas ilr", cex.main=1,
       xlim=c(min(G2[,1],H2[,1],G2[,2],H2[,2],0)*1.1,max(G2[,1],H2[,1],G2[,2],H2[,2],0)*1.2),
       ylim=c(min(G2[,1],H2[,1],G2[,2],H2[,2],0)*1.1,max(G2[,1],H2[,1],G2[,2],H2[,2],0)*1.2),
       xlab=paste("(f) CP1 (",(round(100*b.x2.ilr$importance[2,1],digits=1)), " % )"),
       ylab=paste(" CP2 (", (round(100*b.x2.ilr$importance[2,2],digits=1)), " % )"),type="n")
       text(G2[,1],G2[,2],lab=rownames(x),col=cor)
       #      REPRESENTAÇÃO DAS SETAS
for(i in 1:4){
       arrows(0,0,H2[i,1],H2[i,2],col="red", length=0.1)
}
text(H2[,1]*1.3,H2[,2]*1.3,lab=colnames(x2),cex=0.9,col="red")

#=============================================================================
#      1.3. FREQUÊNCIAS DE BASES NA 3ª POSIÇÃO
#=============================================================================
       #*************************************************#
       # Figura 5.1.(g) DADOS BRUTOS (VAR. ORIGINAIS)             #
       #*************************************************#
       par(mfrow=c(1,3))
x3=dados[,10:13]    # x2=dados[,6:9];x3=dados[,10:13]; Bases na 3ª posição
       x3=scale(x3, center=TRUE, scale=FALSE)
       s3=svd(x3);pc.x3=prcomp(x3,retx=TRUE,center=TRUE);b.x3=summary(pc.x3) # SVD e PCA
       U3=s3$u; V3=s3$v; D3=diag(s3$d)
G3=sqrt(30)*U3                                          #      X=GH', com G=U e H=VD
H3=sqrt(30)*V3%*%D3
       #      REPRESENTAÇÃO DOS PONTOS
plot(G3[,1],G3[,2], main="Bases na 3ª posição: variáveis orginais", cex.main=1,
       xlim=c(min(G3[,1],H3[,1],G3[,2],H3[,2],0)*1.1,max(G3[,1],H3[,1],G3[,2],H3[,2],0)*1.2),
       ylim=c(min(G3[,1],H3[,1],G3[,2],H3[,2],0)*1.1,max(G3[,1],H3[,1],G3[,2],H3[,2],0)*1.2),
       xlab=paste("(g) CP1 (",(round(100*b.x3$importance[2,1],digits=1)), " % )"),
       ylab=paste(" CP2 (", (round(100*b.x3$importance[2,2],digits=1)), " % )"),
       type="n")
       text(G3[,1],G3[,2],lab=rownames(x),col=cor)
       #      REPRESENTAÇÃO DAS SETAS
for(i in 1:4){
       arrows(0,0,H3[i,1],H3[i,2],col="red", length=0.1)
}
text(H3[,1]*1.22,H3[,2]*1.1,lab=colnames(x3),cex=0.9,col="red")

       #*********************************************************#
       # Figura 5.1.(h) DADOS EM COORDENADAS CLR-TRANSFORMADAS    #
       #*********************************************************#
x3.clr=clr(dados[,10:13]); x3.clr=data.frame(x3.clr)  # Transformação ilr de x1
y3=scale(x3.clr, center=TRUE, scale=FALSE)
s3.clr=svd(y3); pc.x3.clr=prcomp(y3); b.x3.clr=summary(pc.x3.clr)
G3=pc.x3.clr$x%*%diag(sqrt(31-1)/s3.clr$d)
H3=pc.x3.clr$rotation%*%diag(s3.clr$d) #.5    # constante d escala = 2.5
plot(G3[,1],G3[,2], main="Bases na 3ª posição: coordenadas clr", cex.main=1,
       xlim=c(min(G3[,1],H3[,1],G3[,2],H3[,2],0)*1.1,max(G3[,1],H3[,1],G3[,2],H3[,2],0)*1.2),
       ylim=c(min(G3[,1],H3[,1],G3[,2],H3[,2],0)*1.1,max(G3[,1],H3[,1],G3[,2],H3[,2],0)*1.2),
       xlab=paste("(h) CP1 (",(round(100*b.x3.clr$importance[2,1],digits=1)), " % )"),
       ylab=paste(" CP2 (", (round(100*b.x3.clr$importance[2,2],digits=1)), " % )"),type="n")
       text(G3[,1],G3[,2],lab=rownames(x),col=cor)
       #      REPRESENTAÇÃO DAS SETAS
```

```
for(i in 1:4){
      arrows(0,0,H3[i,1],H3[i,2],col="red", length=0.1)
}
text(H3[,1]*1.3,H3[,2]*1.1,lab=colnames(x3),cex=0.9,col="red")

      #*********************************************************#
      # Figura 5.1.(i) DADOS EM COORDENADAS ILR-TRANSFORMADAS    #
      #*********************************************************#
x3.ilr=ilr(dados[,10:13]); z3=data.frame(x3.ilr)      # Transformação ilr de x2
      z3=scale(z3, center=TRUE, scale=FALSE); phi3=ilrBase(dados[,10:13],x3.ilr,4)
      s3.z3=svd(z3); pc.x3.ilr=prcomp(z3); b.x3.ilr=summary(pc.x3.ilr)
      loadz3=phi3%*%pc.x3.ilr$rotation; loadz3=cbind(loadz3,rep(0,4))
      #zzr=pc.x2.ilr$x%*%t(phi2)

G3=sqrt(30)*s3.z3$u                              #      X~= GH'
H3=loadz3%*%diag(s3.clr$d)
plot(G3[,1],G3[,2], main="Bases na 3ª posição: coordenadas ilr", cex.main=1,
      xlim=c(min(G3[,1],H3[,1],G3[,2],H3[,2],0)*1.1,max(G3[,1],H3[,1],G3[,2],H3[,2],0)*1.2),
      ylim=c(min(G3[,1],H3[,1],G3[,2],H3[,2],0)*1.1,max(G3[,1],H3[,1],G3[,2],H3[,2],0)*1.2),
      xlab=paste("(i) CP1 (",(round(100*b.x3.ilr$importance[2,1],digits=1)), " % )"),
      ylab=paste(" CP2 (", (round(100*b.x3.ilr$importance[2,2],digits=1)), " % )"),
      type="n")
      text(G3[,1],G3[,2],lab=rownames(x),col=cor)
      #      REPRESENTAÇÃO DAS SETAS
for(i in 1:4){
      arrows(0,0,H3[i,1],H3[i,2],col="red", length=0.1)
}
text(H3[,1]*1.2,H3[,2]*1.3,lab=colnames(x3),cex=0.9,col="red")

#=============================================================================
#      Figura 5.2. Diagramas ternários
#=============================================================================
xc=dados[,2:13]                  # Dados

dt.x1=acomp(xc,c("A1","T1","C1"))
dt.x2=acomp(xc,c("A3","T3","C3"))
dt.x3=acomp(xc,c("C3","G3","T3"))

par(mfrow=c(1,3))
      plot(dt.x1,cex=1)#0.5)           # Figura 5.2.(a)
      plot(dt.x2,cex=1)#0.5)           # Figura 5.2.(b)
      plot(dt.x3,cex=1)#0.5)           # Figura 5.2.(c)

#############################################################################
#
#      Tabela 5.1. e 5.2. Tabelas dos desvios e correlações para bases em cada
#                         uma das três posições dos codões
#
#############################################################################
x1=dados[,2:5];x2=dados[,6:9]; x3=dados[,10:13];
d.x1=sqrt(diag(var(x1))); d.x2=sqrt(diag(var(x2)));d.x3=sqrt(diag(var(x3)))
cor.x1=cor(x1); cor.x2=cor(x2); cor.x3=cor(x3);
print("Desvios")                        # Tabela 5.1. Desvios
d.x1; d.x2; d.x3
print("Correlações")
cor.x1; cor.x2; cor.x3                  # Tabela 5.2. Correlações

#############################################################################
#      Tabela 5.3. TABELAS VARIAÇÃO DE LOG-RAZÕES
#
#############################################################################
```

```
A=dados[,2:13]
t1=t2=t3=matrix(0,4,4)           # Tabela de variação para 1ª, 2ª e 3ª posição

      #***** tabela de variação de log-razões: 1ª posição ******* #
t1[1,2]=var(log(A[,1]/A[,2]));t1[1,3]=var(log(A[,1]/A[,3]));t1[1,4]=var(log(A[,1]/A[,4]))
t1[2,3]=var(log(A[,2]/A[,3]));t1[2,4]=var(log(A[,2]/A[,4]))
t1[3,4]=var(log(A[,3]/A[,4]))

      #***** tabela de variação de log-razões: 2ª posição ******* #
t2[1,2]=var(log(A[,5]/A[,6]));t2[1,3]=var(log(A[,5]/A[,7]));t2[1,4]=var(log(A[,5]/A[,8]))
t2[2,3]=var(log(A[,6]/A[,7]));t2[2,4]=var(log(A[,6]/A[,8]))
t2[3,4]=var(log(A[,7]/A[,8]))

      #***** tabela de variação de log-razões: 3ª posição ******* #
t3[1,2]=var(log(A[,9]/A[,10]));t3[1,3]=var(log(A[,9]/A[,11]));t3[1,4]=var(log(A[,9]/A[,12]))
t3[2,3]=var(log(A[,10]/A[,11]));t3[2,4]=var(log(A[,10]/A[,12]))
t3[3,4]=var(log(A[,11]/A[,12]))
print("tabelas de variação")

t1; t2;t3                 # Tabela de variação para 1ª, 2ª e 3ª posição

##############################################################################
#     ESTUDO 2: Frequências relativas das bases nas três posições de um condão,
#                de forma conjunta
##############################################################################
#-----------------------------------------------------------------------------
#                   Figura 5.3. BIPLOT CLÁSSICOS
#-----------------------------------------------------------------------------

      #*****************************************************#
      # Figura 5.3. (a) DADOS BRUTOS (VAR. ORIGINAIS)            #
      #*****************************************************#
      par(mfrow=c(1,3))
x=scale(dados[,2:13], center=TRUE, scale=FALSE)
      s=svd(x); pc.x=prcomp(x,retx=TRUE,center=TRUE); b.x=summary(pc.x) # SVD e PCA
      U=s$u; V=s$v; D=diag(s$d)
G=sqrt(31-1)*U                                    #     X=GH', com G=U e H=VD
H=sqrt(31-1)*V%*%D*0.8
      #     REPRESENTAÇÃO DOS PONTOS
plot(G[,1],G[,2], main="Biplot clássico - dados completos: variáveis orginais",
cex.main=0.9,
      xlim=c(min(G[,1],H[,1],G[,2],H[,2],0)*1.1,max(G[,1],H[,1],G[,2],H[,2],0)*1.2),
      ylim=c(min(G[,1],H[,1],G[,2],H[,2],0)*1.1,max(G[,1],H[,1],G[,2],H[,2],0)*1.2),
      xlab=paste("(a) CP1 (",(round(100*b.x$importance[2,1],digits=1)), " % )"),
      ylab=paste(" CP2 (", (round(100*b.x$importance[2,2],digits=1)), " % )"),
      type="n")
      text(G[,1],G[,2],lab=rownames(x),col=cor)
      #     REPRESENTAÇÃO DAS SETAS
for(i in 1:12){
      arrows(0,0,H[i,1],H[i,2],col="red", length=0.1)
}
text(H[,1]*1.12,H[,2]*1.1,lab=colnames(x),cex=0.9,col="red")

      #*********************************************************#
      # Figura 5.3. (b) DADOS EM COORDENADAS CLR-TRANSFORMADAS   #
      #*********************************************************#
x.clr=clr(dados[,2:13]); x.clr=data.frame(x.clr)     # Transformação ilr de x1
y=scale(x.clr, center=TRUE, scale=FALSE)
s.clr=svd(y); pc.x.clr=prcomp(y); b.x.clr=summary(pc.x.clr)
G=pc.x.clr$x%*%diag(sqrt(31-1)/s.clr$d)
```

```
H=pc.x.clr$rotation%*%diag(s.clr$d)*0.9# constante d escala = 2.5
plot(G[,1],G[,2], main="Biplot clássico - dados completos: coordenadas clr", cex.main=0.9,
       xlim=c(min(G[,1],H[,1],G[,2],H[,2],0)*1.1,max(G[,1],H[,1],G[,2],H[,2],0)*1.2),
       ylim=c(min(G[,1],H[,1],G[,2],H[,2],0)*1.1,max(G[,1],H[,1],G[,2],H[,2],0)*1.2),
       xlab=paste("(b) CP1 (",(round(100*b.x.clr$importance[2,1],digits=1)), " % )"),
       ylab=paste(" CP2 (", (round(100*b.x.clr$importance[2,2],digits=1)), " % )"),
       type="n")
       text(G[,1],G[,2],lab=rownames(x),col=cor)
       #      REPRESENTAÇÃO DAS SETAS
for(i in 1:12){
       arrows(0,0,H[i,1],H[i,2],col="red", length=0.1)
}
text(H[,1]*1.12,H[,2]*1.1,lab=colnames(x),cex=0.9,col="red")

       #*****************************************************************#
       #      Figura 5.3. (c) DADOS EM COORDENADAS ILR-TRANSFORMADAS     #
       #*****************************************************************#
x.ilr=ilr(dados[,2:13]); z=data.frame(x.ilr)          # Transformação ilr de x
       z=scale(z, center=TRUE, scale=FALSE); phi=ilrBase(dados[,2:13],x.ilr,12)
       sz=svd(z); pc.x.ilr=prcomp(z); b.x.ilr=summary(pc.x.ilr)
       loadz=phi%*%pc.x.ilr$rotation; loadz=cbind(loadz,rep(0,12))
       #zzr=pc.x2.ilr$x%*%t(phi2)

G=sqrt(30)*sz$u                                 #      X~= GH'
H=loadz%*%diag(s.clr$d)*0.9
       #      REPRESENTAÇÃO DOS PONTOS
plot(G[,1],G[,2], main="Biplot clássico - dados completos: coordenadas ilr", cex.main=0.9,
       xlim=c(min(G[,1],H[,1],G[,2],H[,2],0)*1.1,max(G[,1],H[,1],G[,2],H[,2],0)*1.2),
       ylim=c(min(G[,1],H[,1],G[,2],H[,2],0)*1.1,max(G[,1],H[,1],G[,2],H[,2],0)*1.2),
       xlab=paste("(c) CP1 (",(round(100*b.x.ilr$importance[2,1],digits=1)), " % )"),
       ylab=paste(" CP2 (", (round(100*b.x.ilr$importance[2,2],digits=1)), " % )"),type="n")
       text(G[,1],G[,2],lab=rownames(x),col=cor)
       #      REPRESENTAÇÃO DAS SETAS
for(i in 1:12){
       arrows(0,0,H[i,1],H[i,2],col="red", length=0.1)
}
text(H[,1]*1.12,H[,2]*1.3,lab=colnames(x),cex=0.9,col="red")

##############################################################################
#      Tabela 5.4. e 5.5. DESVIOS E DE CORRELAÇÕES NAS 3 POSIÇÕES DOS CODÕES
#
##############################################################################
x=dados[,2:13];d.x=sqrt(diag(var(x))); cor.x=cor(x);
print("Desvio")
d.x                              # Tabela 5.4. Desvios
print("Correlações")
cor.x                            # tabela 5.5. Correlações

##############################################################################
#      Figura 5.4. Diagramas ternários para composição completa
#
#-----------------------------------------------------------------------------
xc=dados[,2:13]     #,center=TRUE,scale=FALSE)# Dados centrados

dt.xa=acomp(xc,c("A1","A2","C3"))
dt.xb=acomp(xc,c("A1","A2","G3"))
dt.xc=acomp(xc,c("C1","C2","T3"))

par(mfrow=c(1,3))
       plot(dt.xa,cex=1)#0.5)           # Figura 5.4. (a)
       plot(dt.xb,cex=1)#0.5)           # Figura 5.4. (b)
```

```
       plot(dt.xc,cex=1)#0.5)           # Figura 5.4. (c)

##############################################################################
#      Figura 5.5. Biplot robusto referente às bases nas três posições dos codões
#
#=============================================================================
#      Figura 5.5. (Esquerda) Variáveis originais
#*****************************************************************************
#       1° PASSO:   Estimativa robusta de sigma e     svd(sigma)=GLG'
#-----------------------------------------------------------------------------
       par(mfrow=c(1,2))
x=dados[,2:13]
#x=scale(x, center=TRUE)
rob.est=covMcd(x,alpha=0.7)      # Estimativa robusta de sigma e mu
                    # http://127.0.0.1:22225/library/robustbase/html/covMcd.html
       sigma=rob.est$cov         # Estimativa robusta da matriz de covariâncias
       mu=rob.est$center         # Estimativa robusta do vetor das médias
s=svd(sigma)              # svd(sigma)=GLG'
       G=s$v; L=s$d        # Matrizes G' e L
       #sum(diag(sigma)); sum(L)
       rownames(G,do.NULL=TRUE,prefix="col")
       rownames(G)=colnames(x)
       #-------------------------------------------------------------------#
       # 2° PASSO:  Determinação de damatriz de scores X*=(X-U)G e biplot #
       #-------------------------------------------------------------------#
um=rep(1,31)                     # Vetor de entradas unitárias 31x1
x.scores=(as.matrix(x-um%*%t(mu)))%*%(as.matrix(G))  # X*=(X-U)G
       F=x.scores%*%diag(sqrt(1/L))            # scores
       H=G%*%diag(sqrt(L))*15           # H=DV de svd(x)=UDV
       #      REPRESENTAÇÃO DOS PONTOS
plot(F[,1],F[,2], main="Biplot robusto - dados completos: variáveis originais", cex.main=0.8,
       xlim=c(min(F[,1],H[,1],F[,2],H[,2],0)*1.1,max(F[,1],H[,1],F[,2],H[,2],0)*1.2),
       ylim=c(min(F[,1],H[,1],F[,2],H[,2],0)*1.1,max(F[,1],H[,1],F[,2],H[,2],0)*1.2),
       xlab=paste(" PC 1  (", (round(100*L[1]/sum(L),digits=1)), " % )"),
       ylab=paste(" PC 2  (", (round(100*L[2]/sum(L),digits=1)), " % )"), type="n")
       #      REPRESENTAÇÃO DAS SETAS
text(F[,1],F[,2],lab=rownames(x),col=cor)

for(i in 1:12){
       arrows(0,0,H[i,1],H[i,2],col="red", length=0.1)
}
text(H[,1]*1.12, H[,2],lab=colnames(x),cex=0.9,col="red")

##############################################################################
#      Figura 5.5. (Direita) Biplot robusto em coordenadas ilr
#                   com identificação de outliers
#-----------------------------------------------------------------------------
#  ******* construção biplot com recurso ao Package mvoutilier  *********
##############################################################################

res=mvoutlier.CoDa(dados[,2:13])
       plot(res,which="biplot",onlyout=FALSE,symb=TRUE,symbtxt=TRUE)

##############################################################################
#
#      Figura 5.6. Diagramas ternários para composição completa
#
#-----------------------------------------------------------------------------
xc=dados[,2:13]     #,center=TRUE,scale=FALSE)# Dados centrados

dt.xa=acomp(xc,c("C2","C3","A3"))
```

```
dt.xb=acomp(xc,c("A1","A2","C1"))
dt.xc=acomp(xc,c("T2","T3","C1"))

par(mfrow=c(1,3))
      plot(dt.xa,cex=1)#0.5)          # Figura 5.6. (a)
      plot(dt.xb,cex=1)#0.5)          # Figura 5.6. (b)
      plot(dt.xc,cex=1)#0.5)          # Figura 5.6. (c)

#############################################################################
#
#     Estudo 3: Análise dados fundidos - soma das frequências de cada
#               uma das bases
#============================================================================
require(RODBC)                        # Package com funções para a conexão
ficheiro=odbcConnectExcel("CodonSpaceVersionSPE.xls")
dados=sqlFetch(ficheiro, sqtable="Fusão")
x.fusao=dados[,2:5]
head(x.fusao)
cor=dados[,7]# 1: preto, 2: vermelho, 3: verde, 4: azul, 6: Magenta
library(compositions)
#============================================================================
#     Figura 5.7. AMALGAMAÇÃO - BIPLOT CLÁSSICO
#============================================================================
      #*****************************************************#
      # Figura 5.7.(a) DADOS BRUTOS (VAR. ORIGINAIS)              #
      #*****************************************************#
      par(mfrow=c(1,2))
      x.fusao=scale(x.fusao, center=TRUE, scale=FALSE)
      s.fusao=svd(x.fusao); pc.fusao=prcomp(x.fusao,retx=TRUE,center=TRUE)
               b.fusao=summary(pc.fusao)      # SVD e PCA
      U.fusao=s.fusao$u; V.fusao=s.fusao$v; D.fusao=diag(s.fusao$d)
G=sqrt(30)*U.fusao                                    #     X=GH', com G=U e H=VD
H=sqrt(30)*V.fusao%*%D.fusao*1.6
      rownames(H,do.NULL=TRUE,prefix="col")   # Nomes das variáveis
      rownames(H)=colnames(x.fusao)

#            REPRESENTAÇÃO DOS PONTOS
plot(G[,1],G[,2], main="(a) Fusão: Biplot para coordenadas originais", cex.main=1,
      xlim=c(min(G[,1],H[,1],G[,2],H[,2],0)*1.1,max(G[,1],H[,1],G[,2],H[,2],0)*1.2),
      ylim=c(min(G[,1],H[,1],G[,2],H[,2],0)*1.1,max(G[,1],H[,1],G[,2],H[,2],0)*1.2),
      xlab=paste(" CP1 (",(round(100*b.fusao$importance[2,1],digits=1)), " % )"),
      ylab=paste(" CP2 (", (round(100*b.fusao$importance[2,2],digits=1)), " % )"),type="n")
      text(G[,1],G[,2],lab=rownames(x.fusao),col=cor)
      #     REPRESENTAÇÃO DAS SETAS
for(i in 1:4){
      arrows(0,0,H[i,1],H[i,2],col="red", length=0.1)
}
text(H[,1]*1.15,H[,2]*1.1,lab=colnames(x.fusao),cex=0.9,col="red")

      #*********************************************************#
      # Figura 5.7.(b) DADOS EM COORDENADAS CLR-TRANSFORMADAS    #
      #*********************************************************#
clr.fusao=clr(dados[,2:5]); clr.fusao=data.frame(clr.fusao)  # Transformação clr
      Y=scale(clr.fusao, center=TRUE, scale=FALSE)
      svd.Y=svd(Y); pc.Y=prcomp(Y); b.Y=summary(pc.Y)
G=pc.Y$x%*%diag(sqrt(31-1)/svd.Y$d)     # *5
H=pc.Y$rotation%*%diag(svd.Y$d)*2

      #     REPRESENTAÇÃO DOS PONTOS
plot(G[,1],G[,2], main="(b) Fusão: Biplot composicional", cex.main=1,
      xlim=c(min(G[,1],H[,1],G[,2],H[,2],0)*1.1,max(G[,1],H[,1],G[,2],H[,2],0)*1.2),
```

```
       ylim=c(min(G[,1],H[,1],G[,2],H[,2],0)*1.1,max(G[,1],H[,1],G[,2],H[,2],0)*1.2),
       xlab=paste(" CP1 (",(round(100*b.Y$importance[2,1],digits=1)), " % )"),
       ylab=paste(" CP2 (", (round(100*b.Y$importance[2,2],digits=1)), " % )"),type="n")
       text(G[,1],G[,2],lab=rownames(x.fusao),col=cor)

       #      REPRESENTAÇÃO DAS SETAS
for(i in 1:4){
       arrows(0,0,H[i,1],H[i,2],col="red", length=0.1)
}
text(H[,1]*1.15,H[,2]*1.1,lab=colnames(Y),cex=0.9,col="red")

##############################################################################
#
#      Estudo 4: Análise de dados fundidos em termos de C+G e  A+T
#
#=============================================================================
require(RODBC)                          # Package com funções para a conexão
ficheiro=odbcConnectExcel("CodonSpaceVersionSPE.xls")
dados=sqlFetch(ficheiro, sqtable="FusãoCG")
x.CG=dados[,2:7]
head(x.CG)
cor=dados[,9]# 1: preto, 2: vermelho, 3: verde, 4: azul, 6: Magenta
library(compositions)
#=====================================================================
#      Figura 5.8.  BIPLOT CLÁSSICO
#=====================================================================
       #*****************************************************#
       # Figura 5.8.(a) DADOS BRUTOS (VAR. ORIGINAIS)              #
       #*****************************************************#
       par(mfrow=c(1,2))
       x.CG=scale(x.CG, center=TRUE, scale=FALSE)
       s.CG=svd(x.CG); pc.CG=prcomp(x.CG,retx=TRUE,center=TRUE)
                b.CG=summary(pc.CG)                                  # SVD e PCA
       U.CG=s.CG$u; V.CG=s.CG$v; D.CG=diag(s.CG$d)
G=sqrt(30)*U.CG                                      #      X=GH', com G=U e H=VD
H=sqrt(30)*V.CG%*%D.CG
       rownames(H,do.NULL=TRUE,prefix="col")   # Nomes das variáveis
       rownames(H)=colnames(x.CG)

#             REPRESENTAÇÃO DOS PONTOS
plot(G[,1],G[,2], main="(a) Fusão pares CG e AT: Coordenadas originais", cex.main=0.9,
       xlim=c(min(G[,1],H[,1],G[,2],H[,2],0)*1.1,max(G[,1],H[,1],G[,2],H[,2],0)*1.2),
       ylim=c(min(G[,1],H[,1],G[,2],H[,2],0)*1.1,max(G[,1],H[,1],G[,2],H[,2],0)*1.2),
       xlab=paste(" CP1 (",(round(100*b.CG$importance[2,1],digits=1)), " % )"),
       ylab=paste(" CP2 (", (round(100*b.CG$importance[2,2],digits=1)), " % )"),type="n")
       text(G[,1],G[,2],lab=rownames(x.CG),col=cor)

       #      REPRESENTAÇÃO DAS SETAS
for(i in 1:6){
       arrows(0,0,H[i,1],H[i,2],col="red", length=0.1)
}
text(H[,1]*1.12,H[,2]*1.1,lab=colnames(x.CG),cex=0.9,col="red")

       #*********************************************************#
       # Figura 5.8.(b) DADOS EM COORDENADAS CLR-TRANSFORMADAS    #
       #*********************************************************#
clr.CG=clr(dados[,2:7]); clr.CG=data.frame(clr.CG)    # Transformação clr de x1
       Y=scale(clr.CG, center=TRUE, scale=FALSE)
       svd.Y=svd(Y); pc.Y=prcomp(Y); b.Y=summary(pc.Y)
G=pc.Y$x%*%diag(sqrt(31-1)/svd.Y$d)     # *5
H=pc.Y$rotation%*%diag(svd.Y$d)*1.6
```

```
       #      REPRESENTAÇÃO DOS PONTOS
plot(G[,1],G[,2], main="(b) Fusão dos pares CG e AT: Biplot composicional",cex.main=0.9,
       xlim=c(min(G[,1],H[,1],G[,2],H[,2],0)*1.1,max(G[,1],H[,1],G[,2],H[,2],0)*1.2),
       ylim=c(min(G[,1],H[,1],G[,2],H[,2],0)*1.1,max(G[,1],H[,1],G[,2],H[,2],0)*1.2),
       xlab=paste(" CP1 (",(round(100*b.Y$importance[2,1],digits=1)), " % )"),
       ylab=paste(" CP2 (", (round(100*b.Y$importance[2,2],digits=1)), " % )"), type="n")
       text(G[,1],G[,2],lab=rownames(x.CG),col=cor)
       #      REPRESENTAÇÃO DAS SETAS
for(i in 1:6){
       arrows(0,0,H[i,1],H[i,2],col="red", length=0.1)
}
text(H[,1]*1.15,H[,2]*1.1,lab=colnames(x.CG),cex=0.9,col="red")

##############################################################################
#      Tabela 5.6. Tabela de correlações de dados fundidos pela soma A+T e C+G
#
#-----------------------------------------------------------------------------
cor(x.CG)           # Correlação entre AT e CG

##############################################################################
#      Tabela 5.7. tabela de variação de log-razões: amalgamação AT e CG
==============================================================================
A=dados[,2:7]
t1=matrix(0,6,6)                # Tabela de variação

       #*********************************************************#
       #            Triângulo superior: variâncias               #
       #*********************************************************#
t1[1,2]=var(log(A[,1]/A[,2]));t1[1,3]=var(log(A[,1]/A[,3]));
       t1[1,4]=var(log(A[,1]/A[,4])); t1[1,5]=var(log(A[,1]/A[,5]));
       t1[1,6]=var(log(A[,1]/A[,6]));
t1[2,3]=var(log(A[,2]/A[,3]));t1[2,4]=var(log(A[,2]/A[,4]));
       t1[2,5]=var(log(A[,2]/A[,5]));t1[2,6]=var(log(A[,2]/A[,6]));
t1[3,4]=var(log(A[,3]/A[,4])); t1[3,5]=var(log(A[,3]/A[,5]));
        t1[3,6]=var(log(A[,3]/A[,6]));
t1[4,5]=var(log(A[,4]/A[,5]));t1[4,6]=var(log(A[,4]/A[,6]));
t1[5,6]=var(log(A[,5]/A[,6]))

t1           # Tabela 5.7.

##################################### FIM ####################################
```